Ministry of Education and Science of Ukraine

Sumy State University

Educational and Scientific Institute of Business, Economics and Management

Department of Economic Cybernetics

BACHELOR'S QUALIFICATION WORK

on the topic "Assessment of the fraud probability in the process of lending to the bank customers"

Completed student of 4th course, group AB-71-8a.en
(course number) (group code)

Specialties 051 "Economics" (Business analytics)

V.V. Radko
(student's last name, initials)

Supervisor PhD in Economics, docent,
associate professor H.M. Yarovenko
(position, degree, last name, initials)

Sumy-2021

Ministry of Education and Science of Ukraine
Sumy State University
Educational and Scientific Institute of Business, Economics and Management
Department of Economic Cybernetics

APPROVE
Head of the Department
Dr. Econ. Sciences, Professor
____________ Kuzmenko OV
"__"__________________2021

TASK
FOR THE BACHELOR'S QUALIFICATION WORK
in the direction of training 051 Economic (Business analytics)
student 4th year of the group АБ-71-8а.ан

Radko Viktoriia Viktorivna

1. Topic of the work: Assessment of the probability of fraud in the process of lending to bank customers approved by order of the university 0382-III from 15.03.2021.
2. The deadline for the student to submit the completed work "_ " _______ 2021.
3. The purpose of the work is to analyze the criminogenic situation associated with fraud in the process of lending to bank customers and to build mathematical models to determine the probability of their occurrence using the Python programming language.
4.The object of the study is fraudulent transactions related to customer lending and personal data of bank customers.
5. The subject of research is mathematical models and analytical tools, such as the Python programming language.
6. Thesis is performed on materials of JSC CB 'PrivatBank'.
7. Indicative plan of qualification work, terms of submission of sections to the head and the maintenance of tasks for performance of the set purpose
Chapter 1 Theoretical principles of financial fraud in the process of lending to bank customers____
In chapter 1 to reveal the essence of the object of research - fraudulent transactions related to lending to bank customers, analyze existing approaches to modeling and analyzing banking operations to identify signs of fraud and develop a conceptual model for estimating the likelihood of fraud_______________________________

Chapter 2 Mathematical model for detecting the probability of fraud in the lending process____
In chapter 2 to describe the input data for modeling: statistical analysis and visualization, describe a mathematical model: stages of constructio_____________

Chapter 3 Modeling the process of identifying bank customers for credit fraud____
In chapter 3 to implement mathematical models using Python, check the adequacy, accuracy and quality of the model and compare them________________________

8. Consultations on work:

Chapter	Consultant	Signature, data	
		Task issued by	Task accepted by
1		Yarovenko H.M.	Radko V.V.
2		Yarovenko H.M.	Radko V.V.
3		Yarovenko H.M.	Radko V.V..

9. Date of issue of the task“___”__________20__ p.

Supervisor ___________ _H.M. Yarovenko__

Signuture Initials, surname

Received the task to perform ___________ _V.V.Radko____

Signuture Initials, surname

ABSTRACT
of the qualifying work
for obtaining the educational and qualification level "bachelor"

Radko Victoria Viktorivna
(surname, name, patronymic of the student)

Assessment of the probability of fraud in the process of lending to bank customers

The relevance of the topic chosen for the study is determined by the fact that in recent years the percentage of growth of financial fraud in lending to bank customers is growing rapidly and every year there are new schemes of credit fraud. Therefore, this issue should be given more attention and apply more serious, so mathematical approaches will assess the likelihood of fraud in the process of lending to bank customers.

The purpose of the qualification work is to analyze the criminogenic situation associated with fraud in the process of lending to bank customers and build mathematical models to determine the probability of their occurrence using the Python programming language.

The object of research is fraudulent transactions related to customer lending and personal data of bank customers.

The subject of research is mathematical models and analytical tools, such as the Python programming language.

The objectives of the study are:

1) to reveal the essence of the study - fraudulent credit operations;

2) to analyze existing approaches to modeling and analysis of banking operations to identify signs of fraud;

3) to develop a conceptual model for estimating the likelihood of fraud;

4) to describe the input data for modeling and perform statistical data analysis and visualization;

5) to build mathematical models and interpret the obtained simulation results.

Research methods: deduction, analysis, comparison, modeling.

The information base of the qualification work is analytical reviews of scientific publications of domestic and foreign authors.

The main scientific result of the qualification work is as follows: revealed the essence of fraudulent transactions in the process of lending to bank customers, analyzed existing methods and approaches to modeling banking operations. signs of fraud and mathematical models are built to assess the probability of credit fraud by influencing the individual performance of the bank's customers on the result.

The results can be used by modern commercial banks in the process of lending to customers.

Keywords: credit fraud, credit risk, scoring, Python, logistic regression, decision tree, neural network, ROC-curve.

The content of the qualification work is presented on 68 pages. The list of the used sources from 40 names, placed on 4 pages. The work contains 6 tables, 56 figures.

Year of performance of qualification work - 2021.

Year of protection of work – 2021.

PLAN

INTRODUCTION

The rapid development of globalization of financial processes and the involvement of information technology contribute to the fact that the financial and banking system becomes more vulnerable to fraud.

The financial market of Ukraine is no exception in the above processes. After all, in recent years, the financial structure of Ukraine is gradually undergoing positive changes, which are aimed at defining our state in the world economic arena.

The economic crisis, low incomes, the creation of a number of commercial banks and specialized financial companies and the development of information technology have created the basis for the spread of the phenomenon as "banking fraud".

The problem of financial fraud in lending to bank customers is very common in the world.

Banking fraud is a type of fraud, the process of which is hidden and which systematically affects the financial security of the bank, leads to financial losses, loss of trust and reputation among customers, and can cause total bankruptcy of the banking institution.

The relevance of this study is determined by the fact that in recent years the growth rate of financial fraud in lending to bank customers is growing rapidly and every year there are new schemes of credit fraud. Therefore, this issue should be given more attention and apply more serious, namely mathematical approaches that will assess the likelihood of fraud in the process of lending to bank customers.

Recently, the percentage of credit fraud is growing. In 2020, credit card fraud ranks second, accounting for 29.7% of the five most common financial crimes.

To prevent credit fraud, it is advisable to use mathematical models, by calculating which could identify the transaction and the subject of fraud or a potential customer of the bank, which can commit it. The use of the latest technologies and programming languages simplifies the calculation of such models and allows you to build a model of any level of complexity. Therefore, assessing the likelihood of fraud in the process of lending to bank customers is a topical issue today and practically significant.

The purpose of the thesis is to analyze the criminal situation associated with the issuance of bank loans and the construction of mathematical models to determine the likelihood of fraud and their implementation using the Python programming language.

The object of this study is fraudulent transactions related to lending to bank customers and personal data of bank customers, which will help identify the relationship with the likelihood of credit fraud.

The subject of the research is statistical and analytical tools such as the Python programming language and mathematical models for detecting the impact of a client's personal performance on the likelihood of financial fraud.

When writing the thesis, the following tasks were set:

1) to reveal the essence of the object of research - fraudulent transactions related to lending to bank customers;

2) to analyze existing approaches to modeling and analysis of banking operations to identify signs of fraud;

3) to develop a conceptual model for estimating the probability of fraud;

4) to describe the input data for modeling;

5) to conduct statistical analysis and visualization of data;

6) to describe the mathematical model: stages of construction and its implementation using Python;

7) to interpret the obtained simulation results.

CHAPTER 1 THEORETICAL PRINCIPLES OF FINANCIAL FRAUD IN THE PROCESS OF LENDING TO BANK CUSTOMERS

1.1 The essence of fraudulent transactions related to lending to bank customers

At present, it is safe to say that fraud has the ability to change its form and penetrates into almost all spheres of social life. But the financial sector is the most vulnerable to fraud.

The problem of financial fraud is widespread around the world. Ukraine is no exception. The largest number of crimes, according to the Report to the Nations in 2018 year, is recorded in the banking sector [1].

The banking sector is a branch of the economy that provides the accumulation and distribution of financial and investment resources. The sector also includes the regulation of banking by public authorities, insurance, interaction between producers on consumers of financial services, investment services and credit card [2].

Despite its prevalence, financial fraud does not have a clear and well-established terminological definition. As a result, it has classifications that are more or less related to financial fraud, but sometimes financial fraud in a commercial bank includes other types of financial crimes that are not inherently financial fraud [3].

Banking fraud occurs when someone tries to take money or other assets from a financial institution or from clients of that institution, posing as a bank official [4].

The rapid development of information technology and its introduction into financial structures has led to dozens of new frauds. Criminals use a variety of ways to break the law, but new technologies can circumvent it.

Credit card fraud is a very common method of fraud.

Statistics about top five types of financial theft are shown in Table 1.1 [5].

Table 1.1 - Top Five Types of Financial Theft in 2020 year

Type of identity theft	Number of reports	Percent of total top five
Government benefits applied for/received	394,324	32.0%
Credit card fraud	365,597	29.7%
Miscellaneous identity theft	281,434	22.9%
Business/personal loan	99,667	8.1%
Tax fraud	89,391	7.3%
Total, top five	1,230,413	100.0%

Credit card fraud ranks second (29.7%) among the five most common financial crimes.

The use of payment and credit bank cards, on the one hand, should facilitate payments anywhere in the world and save time. But the spread of the non-cash form of payment puts both the population and the financial structures at risk, as thieves become more inventive.

Recently, forensic economic expertise in the field of financial and credit relations has increasingly recorded fraud involving illegal credit transactions. This is one of the most common methods of financial fraud, which is resorted to by both creditors (banks, funds, associations) and borrowers (businesses, individuals). Thus economic losses of the specified subjects and, first of all, financial and credit establishments, can reach considerable sums in national or foreign currency [6].

Figure 1.1. depicts the growing percentage of illegal credit transactions with credit limit cards [5].

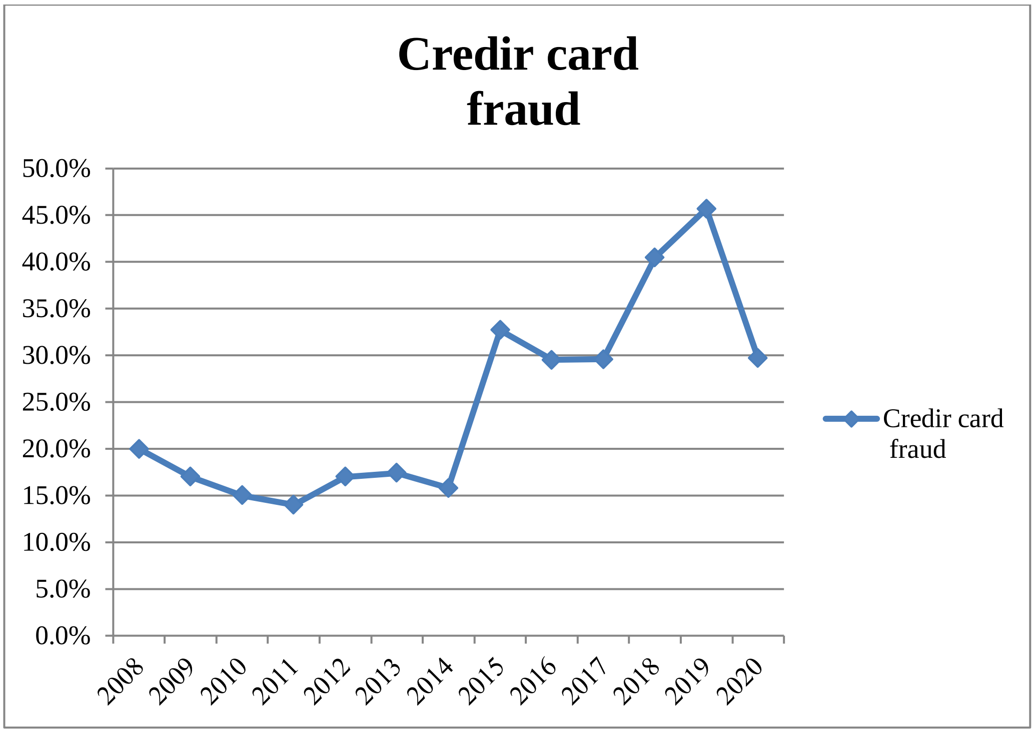


Figure 1.1 – Credit card fraud

There has been an increase in consumer credit recently. And not only their number, but also the share of consumer loans in the total amount of bank loans is growing.

Consumer credit, also called consumer debt, is a loan provided to individuals to purchase goods or services. Most often associated with credit cards, consumer credit also includes other lines of credit, including some loans [7].

Bank`s borrowers who are fraudsters, can deceive bankers mainly in two ways The first of them is to present themselves as another person, have fake documents and after receiving credit funds disappear forever. The second way is to find a frivolous person who is in a difficult financial situation and vouch for her.

The second method is almost non-existent in serious organizations, such as commercial and national banks, most likely such loans are issued in newly established specialized financial companies that provide short-term consumer loans on simplified terms with high interest rates and penalties [8].

Therefore, domestic banks face a big problem when the population does not repay the funds provided.

1.2 Credit risk and its management methods

In a general sense, credit risk is the risk of losses that may result from non-compliance by either party with a financial contract, but this is mainly due to non-payment of the required loan payments [9].

First of all, credit risk assessment is important for any creditor. This is the basis on which the lender can calculate the probability of default by the borrower or the fulfillment of other contractual obligations. In a broader sense, credit risk management is the calculation of the probability that a lender will not receive the proper principal and accrued interest and if this happens, will result in losses and increase the cost of recovering the due debt [10].

Credit risk has its own specific features that every commercial bank must take into account in the process of credit risk management.

First of all, credit risk assessment has many subjective characteristics compared to other financial risks. After all, certain risks, such as changes in exchange rates, interest rate fluctuations or the political situation of cranes are assessed by all banking institutions in the country and not by one bank. Therefore, in the process of providing consumer credit, it is not possible to rely on the generally accepted ideas of other financial institutions, because credit risk is individual, which is associated with a particular borrower. In addition, when assessing the risk of granting a loan to a particular client, not only the financial position of the entity is assessed, but also qualitative indicators such as education, marital status and scope of the client.

Therefore, from the above theoretical point of view, a special methodology should be created and applied to assess a particular credit risk, which would take into account all the features of a particular borrower. Of course, in reality it is quite difficult to implement such an approach, and it is not always appropriate and effective. Thus, the dilemma of credit risk reduction depends significantly on the perfection of its assessment methodology. These methods could be unified and formalized, but each bank makes its own decisions and makes adjustments to such methodologies, because each bank has its own market segment and industry specifics. The methodologies chosen by the bank should take into account all these features, because the characteristics by which some customers are evaluated may be quite unfavorable for others [11].

The main causes of credit risk are:

- inability of the debtor to generate the required amount of money;
- uncertainty of the creditor in the liquidity and value of the borrower's assets;
- moral and ethical qualities of the borrower;
- shortcomings in the concluded and executed credit agreement;
- a risk of the country (economic, political, regulatory) [12].

The Board of the National Bank of Ukraine in 2018 approved the Regulations on the organization of risk management system in banks of Ukraine and banking groups in order to improve the management system in banks of Ukraine risks, taking into account the principles and recommendations of the Basel Committee. Therefore, each bank must create an adequate and effective credit risk management system, which must comply with the following principles:

1) efficiency - ensuring objective assessment and completeness of risk management measures with optimal use of financial resources;

2) timeliness - ensuring timely detection, monitoring, control, reporting and mitigation of all types of risks at all organizational levels ;

3) structure - a clear division of functions, responsibilities and powers of risk management between all departments;

4) separation of responsibilities - separation of control functions from the implementation of bank operations;

5) independence - freedom from circumstances that threaten the impartial performance of the chief risk manager, chief compliance manager, risk management unit and unit of control over compliance with the rules (compliance) of its functions;

6) confidentiality - restriction of access to information that should be protected from unauthorized access,

7) transparency - disclosure by the bank of information on the risk management system and profile risk [13].

Credit risk management is a rather complex process that focuses on the client's ability to repay the loan. The concept of governance is to gather the most detailed information about the borrower. In other words, the bank's management must ensure the optimal balance between profit, liquidity and risk.

The expert method can be upgraded by processing the judgments of experienced banking professionals.

An example of a subjective determination of credit risk is the rating of a borrower's creditworthiness. One of the main approaches to assessing the creditworthiness of the bank's customers is scoring.

Scoring is a classification method, a mathematical model in the form of a weighted sum of certain characteristics, by which the bank, based on previous experience, tries to estimate the probability of loan repayment and interest payments. The scoring system was first used to assess the borrower's credit risk by D. Durand in 1941.

It took into account the following characteristics of the client: age, gender, length of residence in the area, profession, length of service, bank accounts, ownership real estate, the presence of a life insurance policy [14].

Each client takes a survey and provides detailed information, where each indicator and characteristic has its value in points.

Critical evaluation is based on the assumption that people with the same social indicators behave the same. For example, if a bank has already faced a borrower who was undisciplined in the loan repayment process and had certain social characteristics, then the next client with similar indicators may be denied a loan [15].

Depending on the quality of information obtained about the borrower and the method of decision-making, scoring is divided into deductive and empirical.

In particular, in the integrated credit risk management system SAS Credit Scoring for Banking there is an analytical module through which the formation of models for assessing the creditworthiness of borrowers and the possibility of forming algorithms and models for detecting fraud, the ability to perform all types of scoring for different types of tasks and for different types of data.

Fraud Risk Analysis in Application Scoring Fraudsters are divided into three main groups:

– "household" fraudsters are professional fraudsters and borrowers who use the services of professional fraudsters.

Household fraudsters never repay the loan due to financial difficulties. These debtors will not hide from the bank and collectors, and after the court will be forced to return the goods and banks do not receive income from such loans;

– the professional fraudsters constantly change addresses, mobile phones, nowhere officially work;

– The third type is borrowers who attract fraudsters to get loans to start a business and after a while are unable to repay the loan, and eventually become fraudsters themselves.

The losses received by the bank depend on the amount of the loan provided to such fraudsters [16].

After assessing the likelihood of fraud in the process of lending to the borrower, traditional models and scoring cards are used. A scoring card is a model that allows a bank employee to identify the factors that characterize fraudulent intentions and refuse to grant a loan (Table 1.2).

Table 1.2 – Example of scoring card

Indicator	Value	Score
Gender	Male	25
	Female	20
Age	Less then 30 years	30
	30 -35 years	35
	More then 40 years	28

Continuation of table 1.2

Indicator	Value	Score
Education	Secondary education	22
	Secondary special education	28
	Unfinished higher	30
	Higher education	40
	Higher education	40
Marital status	Single	20
	Married	30
Car availability	Yes	40
	No	20

Scoring cards have hundreds of positions that are constantly updated and each bank decides which additional indicator to add.

For a positive decision, the borrower must pass dozens of other checks in addition to the scoring system [15].

Also, to assess the reliability and solvency clients of the bank use models of complex analysis. In addition to quantitative indicators, we must not forget about qualitative analysis. To evaluate quantitative and qualitative indicators allow complex models of complex analysis: the rule of “six c”, CAMPARI, PARTS, PARSER.

American banks in practice apply the rule of “six c”, where all the criteria for selecting customers begin with the letter “c”:

- character (character and reputation of the bank's customer);
- capacity or each flow (ability to repay a loan or cash flow in a timely manner);
- capital (capital, property or the amount of share capital (for legal entities));
- collateral (collateral, types and value of assets);
- conditions (economic situation and its prospects);
- control.

Thus, analysts and managers of American banks make a detailed report not only on the main material characteristics, but also analyze the personal qualities of the borrower, the characteristics of the industry and the scope of the client.

In the English-language literature, seven lending principles are developed and used and are denoted by the following abbreviation CAMPARI:

- C - character - personal qualities of the borrower;
- A - ability - the client's ability to repay the loan; } - M - margin - the expected return of the bank;
- P - purpose - purpose of the loan;
- A-amount - total loan amount;
- R - return - loan repayment terms;
- I - insurance - insurance against the risk of non-repayment of the loan.

Another method of loan requirements is "PARTS", which includes:

- P - purpose - the purpose of the loan;
- A - amount - loan amount;
- R - repayment - repayment of principal and interest);
- T - term - term;
- S - secyrity - collateral, loan collateral

Also, English banks use the customer rating system "PARSER":

- P - Person - information about the potential borrower, his reputation;
- A - Amount - justification of the loan amount;
- R - Repayment - loan repayment options;
- S - Security - security assessment;
- E - Expediency - expediency of the loan;
- R - Remuneration - interest rate on loan risk.

The methods listed in this section credit risk management and borrower reliability assessments are effective and useful, primarily saving time and assessing qualitative and quantitative factors [17].

The advantages and disadvantages of each method are listed in Table 1.3.

Table 1.3 - The advantages and disadvantages of methods credit fraud detection

Type of the model	Advantages	Disadvantages
Classification models (Scoring)	– easy to understand and use; – easy to interpret; – using quality variables; – the method takes into account importance of each variables; – credit process much faster and efficient.	– the method needs constant improvement over time; – personal data can`t be interpreted correctly; – expensive staff training; – the presence of a large array of input data.
Complex models (Rule "six c", CAMPARI, PARTS, PARSER)	– fast and efficient process; – minimal labor costs; – minimum operating bank costs.	– constant updating of information about system grades; – the cost of updating information; – impossibility to implement these methods in small banks (limited information base and funds); – models do not take into account the specific features of borrowers.

Thus, in foreign practice, when considering the reliability, solvency of the client and credit risk assessment comprehensively analyze incomparable indicators, such as economic interests of the bank, profit and human quality, the reputation of the potential borrower.

1.3 Conceptual model for estimating the probability of fraud

The main purpose of building a conceptual model is to identify the likelihood of fraud in the process of lending to bank customers and the future prevention of these threats.

Building a conceptual model is a large-scale process that involves a series of steps, from identifying and analyzing real problems, to building a model to detect signs of credit fraud (Figure 1.2).

This process can be implemented using the following steps:

- problem statement;
- data preparation;

- statistical analysis and visualization;
- model building;
- model forecasting.

Consider the algorithm for modeling the probability of fraud in the process of lending to bank customers:

- determination of input data to build a mathematical model for the detection of credit fraud; , identifying trends and relationships between variables, adjusting data as needed,
- building mathematical models in the form of representing mathematical dependencies that will describe the input data;
- modeling the resulting value and obtaining data that likely to affect the prediction of fraudulent transactions in the lending process,
- interpretation of the obtained simulation results;
- analysing of results and generalization of the trend.

Construction of this model involves the use of statistical data of financial institutions and banks, containing data on bank customers, their gender, marital status, education, place of work, real estate and other personal account balance data.

The choice of these factors is due to the fact that due to little attention is paid to the quality characteristics of the bank's customers

Mobile and Internet banking users are one of the weak links in the banking security system.

This is because the bank is unable to control who the user is. Most often, such operations may contain signs of a cyber threat, ie exposed to a kind of social engineering.

One of the effective ways to combat fraud is to use mathematical methods to build a model that will determine the probability of credit fraud. To solve this problem, having previously considered and investigated various methods, logistic regression, neural networks and decision tree were chosen to build the model.

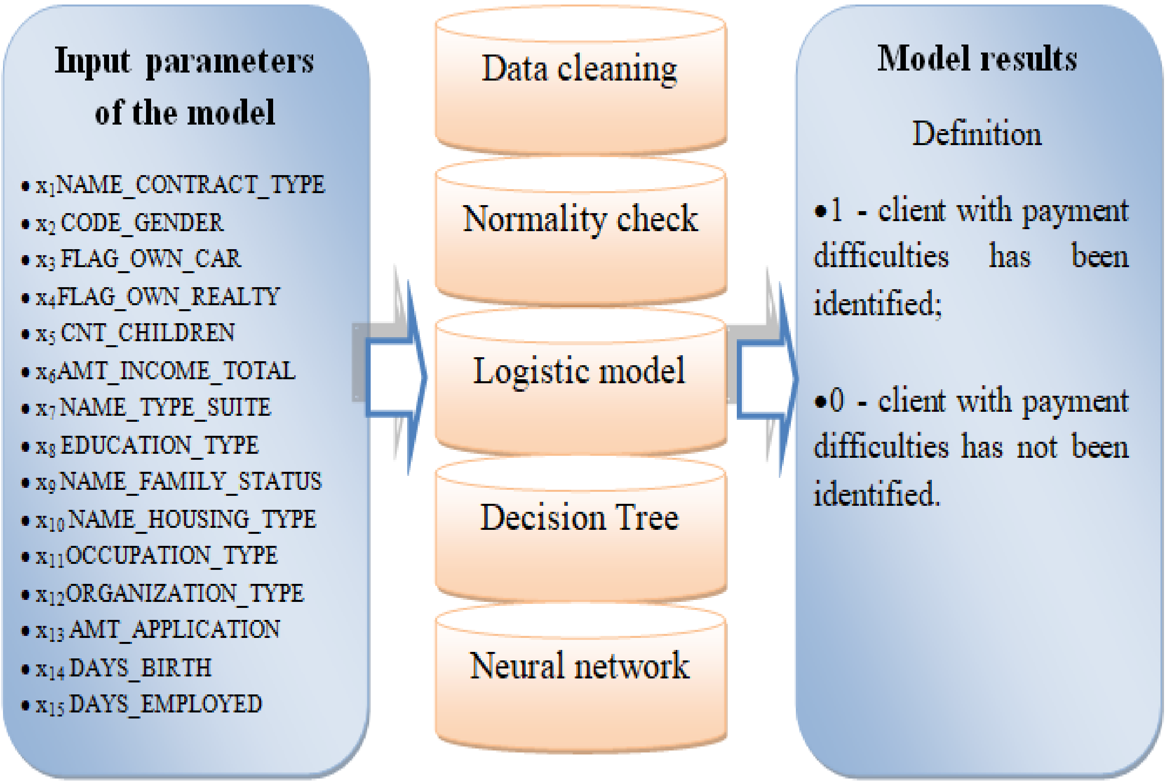

Input parameters of the model
• x_1NAME_CONTRACT_TYPE
• x_2 CODE_GENDER
• x_3 FLAG_OWN_CAR
• x_4FLAG_OWN_REALTY
• x_5 CNT_CHILDREN
• x_6AMT_INCOME_TOTAL
• x_7 NAME_TYPE_SUITE
• x_8 EDUCATION_TYPE
• x_9 NAME_FAMILY_STATUS
• x_{10} NAME_HOUSING_TYPE
• x_{11}OCCUPATION_TYPE
• x_{12}ORGANIZATION_TYPE
• x_{13} AMT_APPLICATION
• x_{14} DAYS_BIRTH
• x_{15} DAYS_EMPLOYED
Data cleaning
Normality check
Logistic model
Decision Tree
Neural network
Model results
Definition
•1 - client with payment difficulties has been identified;
•0 - client with payment difficulties has not been identified.

Figure 1.2 - Conceptual model for detecting signs of fraud in the process of lending to bank customers

In the process of applying the model, the potential threat of fraud during the crediting of the bank's client is checked. If there is such a suspicion, the bank must notify the customer and terminate the transaction.

Based on the selected input data, a conceptual model of fraud detection in the process of lending to bank customers was developed.

Thus, the use of the developed model will prevent typical credit fraud to ensure the security of the banking system.

To implement the economic and mathematical model for detecting the probability of fraud in the process of lending to bank customers programming language Python was chosen.

Python is used not only for programming procedural and object-oriented programming. Python has a lot of libraries for statistics evaluation of the variables, so it can be used for analytics goals.

Thus, the choice of software for building models is quite justified.

CHAPTER 2 MATHEMATICAL MODEL FOR DETECTING THE PROBABILITY OF FRAUD IN THE LENDING PROCESS

2.1 Database analysis

The problem of detecting fraud on credit cards and lending to bank customers includes modeling past transactions for fraud detection. In this thesis, we will build a model that will help manage credit risk and can be used to detect fraud in new credit transactions.

This dataset is loaded to obtain credit card default statistics based on the relevant attributes. The data set contains 307510 rows and 122 columns, including target columns. Here we aim to clear the data set and find some useful information from it.

This study is thematic and aims to give you an idea of the application of EDA in a real business scenario [18].

Research Data Analysis (EDA) is a data analysis approach that uses a variety of methods (mostly graphical) to maximize understanding of the data set:

– to extract important variables;

– to detect deviations and anomalies;

– to check the main assumptions;

– to develop the models [19].

This case study will provide a basic understanding of risk analysis in banking and financial services and understand how the data is used to minimize the risk of losing money when lending to customers.

Also, this section solves and solves the following problems:

– find missing and unnecessary data;

– delete columns with missing and unnecessary data or replace them with appropriate values;

– investigate the data on the normality of the distribution;

– build correlations between the data;

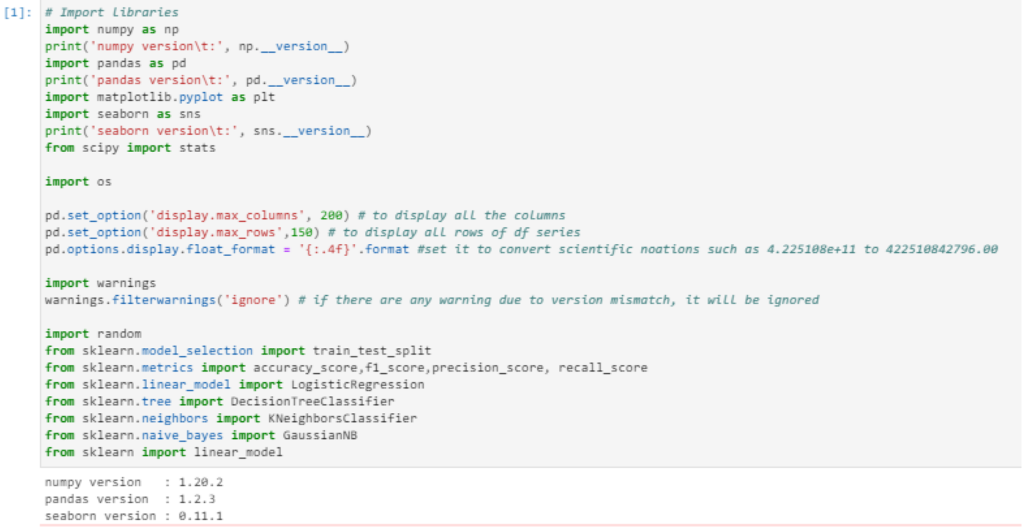
– build mathematical models to identify the relationship between credit fraud and personal data of bank customers;

– analyze the results, assess the adequacy of the model and build a forecast.

2.2 Data preparation, statistical analysis and visualization

To start working with the database, you need to import the necessary Python libraries.

The NumPy package is an indispensable Python companion. It pulls on data analysis, machine learning and scientific computing. Some of the leading Python packages use NumPy as a core element of their infrastructure. Scikit-learn, SciPy, pandas, and tensorflow protect them. In addition to being able to rock the numerical data, the ability to work with NumPy is of great advantage when debugging more complex library scripts (Figure. 2.1).

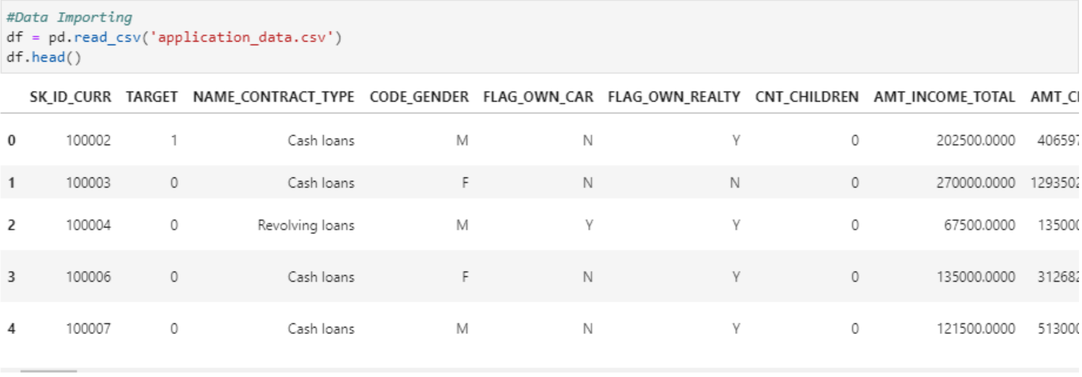


```
[1]: # Import Libraries
import numpy as np
print('numpy version\t:', np.__version__)
import pandas as pd
print('pandas version\t:', pd.__version__)
import matplotlib.pyplot as plt
import seaborn as sns
print('seaborn version\t:', sns.__version__)
from scipy import stats

import os

pd.set_option('display.max_columns', 200) # to display all the columns
pd.set_option('display.max_rows',150) # to display all rows of df series
pd.options.display.float_format = '{:.4f}'.format #set it to convert scientific noations such as 4.225108e+11 to 422510842796.00

import warnings
warnings.filterwarnings('ignore') # if there are any warning due to version mismatch, it will be ignored

import random
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,f1_score,precision_score, recall_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn import linear_model
```

```
numpy version   : 1.20.2
pandas version  : 1.2.3
seaborn version : 0.11.1
```

Figure 2.1 – Importing libraries

All the necessary libraries have been downloaded, now you need to import and read the database (Figure. 2.2):

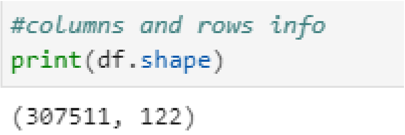


```
#Data Importing
df = pd.read_csv('application_data.csv')
df.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CI
0	100002	1	Cash loans	M	N	Y	0	202500.0000	40659
1	100003	0	Cash loans	F	N	N	0	270000.0000	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0000	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0000	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0000	51300

Figure 2.2 – Importing dataset

After importing the database, you need to perform a series of operations to check the information about existing rows and columns, missing data and delete columns with missing and unnecessary data or replace them with the appropriate values.

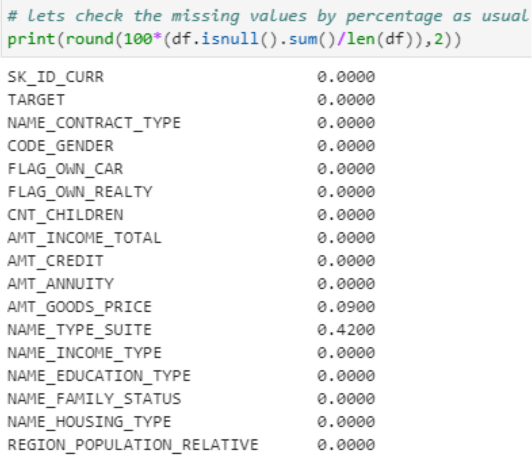


```
#columns and rows info
print(df.shape)
```

```
(307511, 122)
```

Figure 2.3 – Information about columns and rows

This database has 122 variables and 307511 records about different clients of the bank. Find the percentage of missing values of the columns (Figure 2.3):

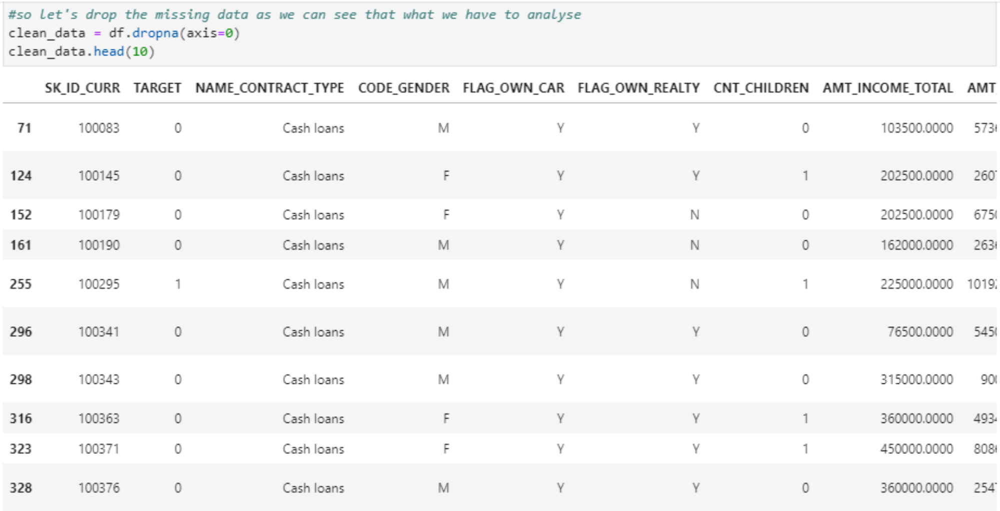


```
# lets check the missing values by percentage as usual
print(round(100*(df.isnull().sum()/len(df)),2))
```

```
SK_ID_CURR                     0.0000
TARGET                         0.0000
NAME_CONTRACT_TYPE             0.0000
CODE_GENDER                    0.0000
FLAG_OWN_CAR                   0.0000
FLAG_OWN_REALTY                0.0000
CNT_CHILDREN                   0.0000
AMT_INCOME_TOTAL               0.0000
AMT_CREDIT                     0.0000
AMT_ANNUITY                    0.0000
AMT_GOODS_PRICE                0.0900
NAME_TYPE_SUITE                0.4200
NAME_INCOME_TYPE               0.0000
NAME_EDUCATION_TYPE            0.0000
NAME_FAMILY_STATUS             0.0000
NAME_HOUSING_TYPE              0.0000
REGION_POPULATION_RELATIVE     0.0000
```

Figure 2.4 – Checking dataset for missing data

Drop all columns from Dataframe for which missing value percentage are more than 50% (Figure. 2.4 - 2.5):

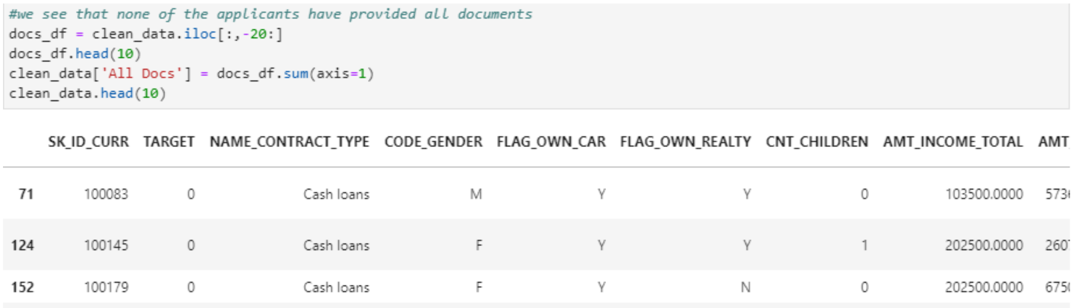


```
#so let's drop the missing data as we can see that what we have to analyse
clean_data = df.dropna(axis=0)
clean_data.head(10)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT
71	100083	0	Cash loans	M	Y	Y	0	103500.0000	573
124	100145	0	Cash loans	F	Y	Y	1	202500.0000	260
152	100179	0	Cash loans	F	Y	N	0	202500.0000	675
161	100190	0	Cash loans	M	Y	N	0	162000.0000	263
255	100295	1	Cash loans	M	Y	N	1	225000.0000	1019
296	100341	0	Cash loans	M	Y	Y	0	76500.0000	545
298	100343	0	Cash loans	M	Y	Y	0	315000.0000	90
316	100363	0	Cash loans	F	Y	Y	1	360000.0000	493
323	100371	0	Cash loans	F	Y	Y	1	450000.0000	808
328	100376	0	Cash loans	M	Y	Y	0	360000.0000	254

Figure 2.5 – Dropping the missing data

Having made a preliminary operation, so that is no client has provided additional information about the document, these columns should be deleted (Figure 2.6):

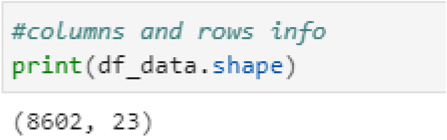


```
#we see that none of the applicants have provided all documents
docs_df = clean_data.iloc[:,-20:]
docs_df.head(10)
clean_data['All Docs'] = docs_df.sum(axis=1)
clean_data.head(10)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT
71	100083	0	Cash loans	M	Y	Y	0	103500.0000	573
124	100145	0	Cash loans	F	Y	Y	1	202500.0000	260
152	100179	0	Cash loans	F	Y	N	0	202500.0000	675

Figure 2.6 – Dropping the document data

After deleting unnecessary and missing information, the database now has 23 variables and 8602 records (Figure. 2.7).

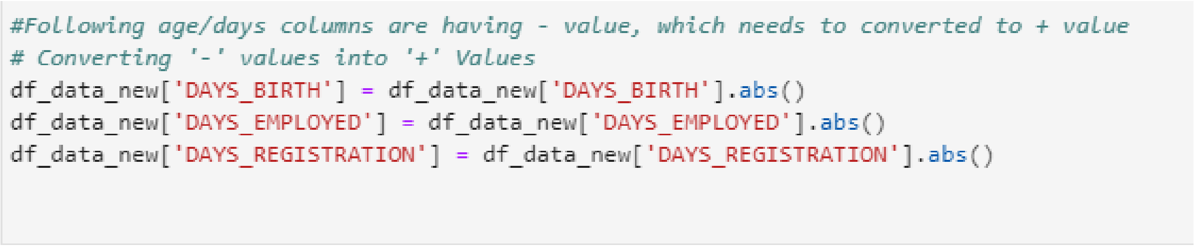


```
#columns and rows info
print(df_data.shape)
```

```
(8602, 23)
```

Figure 2.7 - Information about columns and rows

The database has numerical information that has negative values, in order to obtain correct and clear data you need to convert this information into positive values (Figure 2.8):



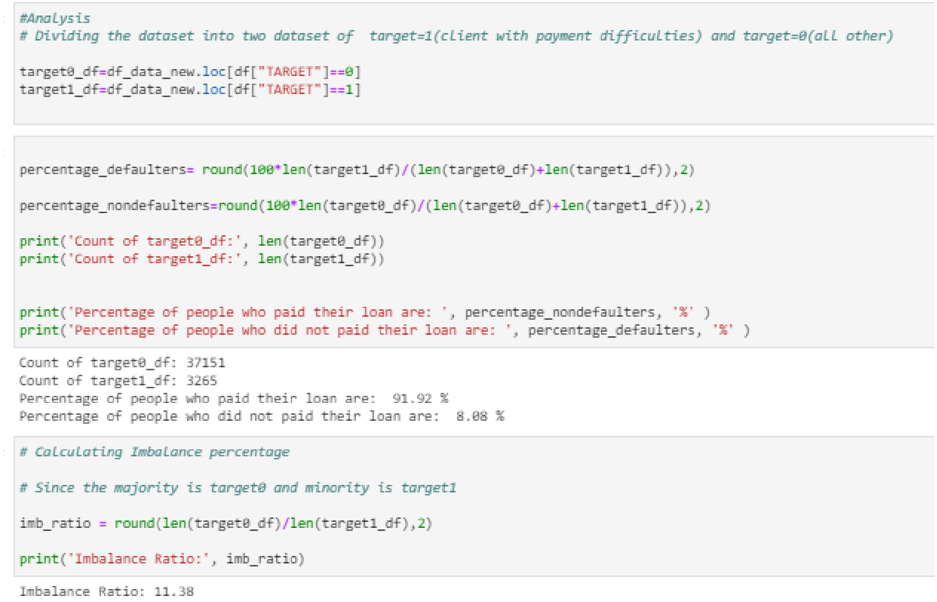
```
#Following age/days columns are having - value, which needs to converted to + value
# Converting '-' values into '+' Values
df_data_new['DAYS_BIRTH'] = df_data_new['DAYS_BIRTH'].abs()
df_data_new['DAYS_EMPLOYED'] = df_data_new['DAYS_EMPLOYED'].abs()
df_data_new['DAYS_REGISTRATION'] = df_data_new['DAYS_REGISTRATION'].abs()
```

Figure 2.8 – Converting negative values to positive

After a series of operations, the database is ready for statistical analysis and data visualization.

To find out which customers have repaid the loan and which credit fraud, you need to divide the database into two parts according to the dependent variable TARGET.

This variable has two values 0 and 1, where 0 - the client has no difficulty in repaying the loan and 1 - the client has difficulty in repaying the loan and the likelihood of fraud.



```
#Analysis
# Dividing the dataset into two dataset of  target=1(client with payment difficulties) and target=0(all other)

target0_df=df_data_new.loc[df["TARGET"]==0]
target1_df=df_data_new.loc[df["TARGET"]==1]
```

```
percentage_defaulters= round(100*len(target1_df)/(len(target0_df)+len(target1_df)),2)

percentage_nondefaulters=round(100*len(target0_df)/(len(target0_df)+len(target1_df)),2)

print('Count of target0_df:', len(target0_df))
print('Count of target1_df:', len(target1_df))


print('Percentage of people who paid their loan are: ', percentage_nondefaulters, '%' )
print('Percentage of people who did not paid their loan are: ', percentage_defaulters, '%' )
```

```
Count of target0_df: 37151
Count of target1_df: 3265
Percentage of people who paid their loan are:  91.92 %
Percentage of people who did not paid their loan are:  8.08 %
```

```
# Calculating Imbalance percentage

# Since the majority is target0 and minority is target1

imb_ratio = round(len(target0_df)/len(target1_df),2)

print('Imbalance Ratio:', imb_ratio)
```

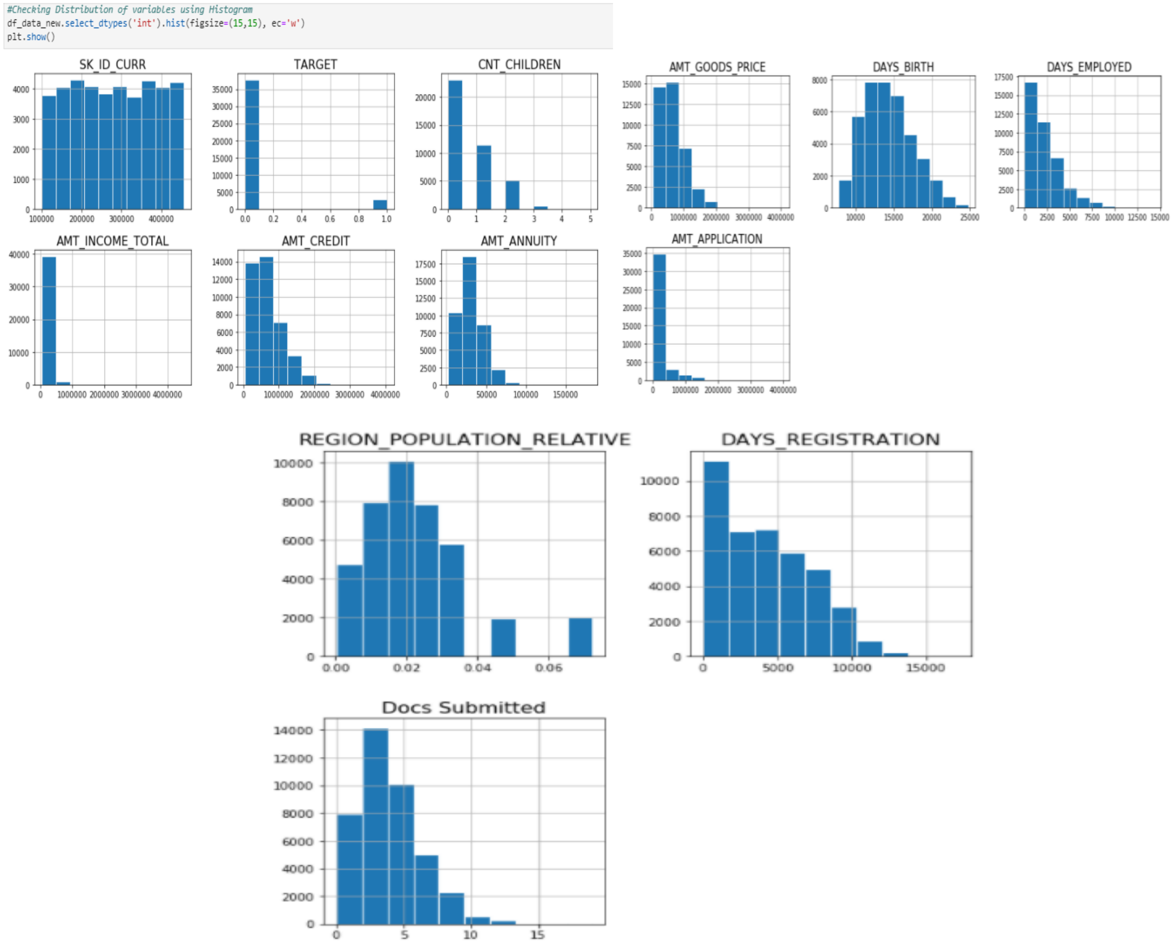
```
Imbalance Ratio: 11.38
```

Figure 2.9 – Imbalance ratio

After dividing the database into two parts according to the dependent variable, we can conclude that 8.08% of customers do not repay the loan to the bank and the imbalance ratio is 11.3 (Figure. 2.9).

It is also necessary to test hypotheses about the nature of data distribution. Hypotheses about distributions are that the distribution in the general population is subject to some specific law. Hypothesis testing is based on a comparison of actual (empirical) frequencies with predicted (theoretical) frequencies to conclude that the actual distribution corresponds to the hypothetical distribution.

Check the data for normality using a histogram (Figure 2.10):



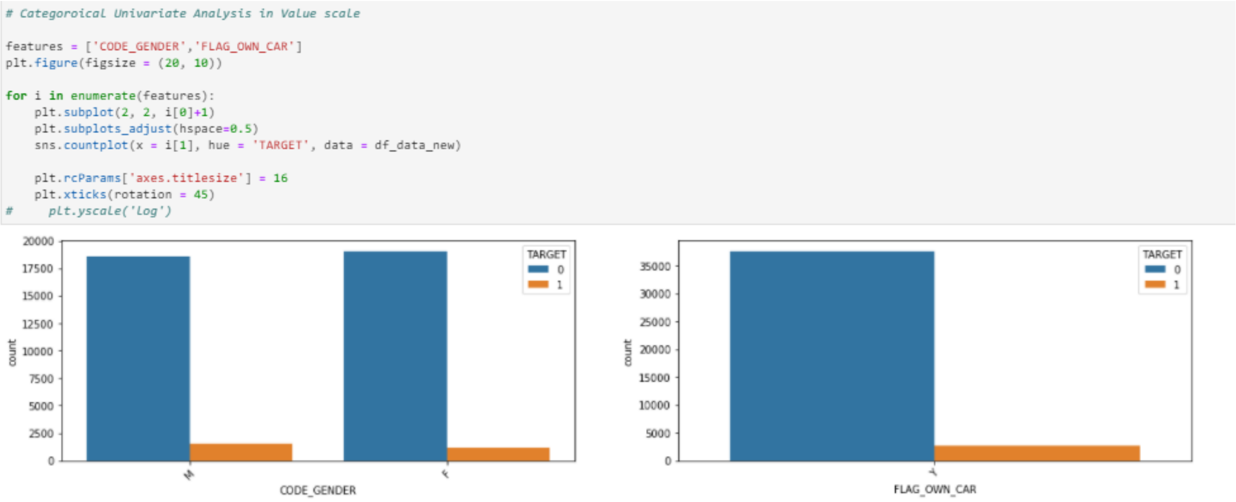
```
#Checking Distribution of variables using Histogram
df_data_new.select_dtypes('int').hist(figsize=(15,15), ec='w')
plt.show()
```

Figure 2.10 - Checking Distribution of variables using Histogram

Data visualization is a visual representation of different types of information using graphs, histograms and charts, which facilitate the representation of large data sets.

Consider visualizing database metrics.

Figure 2.11 shows that the male customer is having the highest count as compare to female customers in both cases.



```
# Categoroical Univariate Analysis in Value scale

features = ['CODE_GENDER','FLAG_OWN_CAR']
plt.figure(figsize = (20, 10))

for i in enumerate(features):
    plt.subplot(2, 2, i[0]+1)
    plt.subplots_adjust(hspace=0.5)
    sns.countplot(x = i[1], hue = 'TARGET', data = df_data_new)

    plt.rcParams['axes.titlesize'] = 16
    plt.xticks(rotation = 45)
#     plt.yscale('log')
```

Figure 2.11 – Distribution of gender depending on target variable

Figure 2.12 shows that the customer having payment difficulties in secondary/ secondary special in both cases.

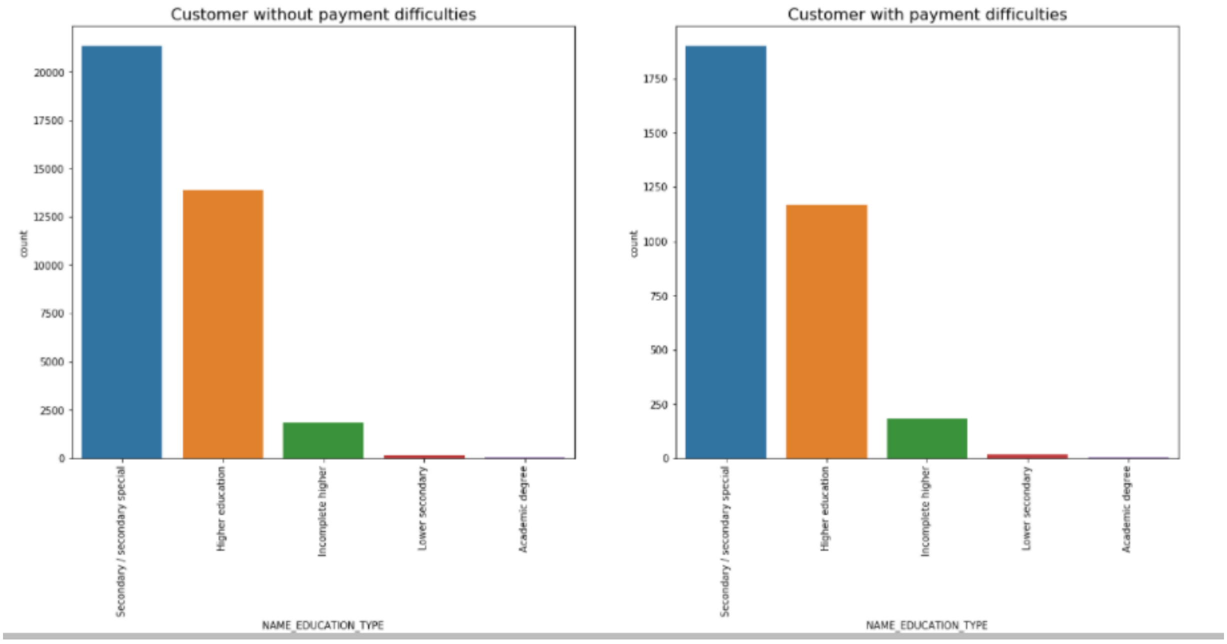
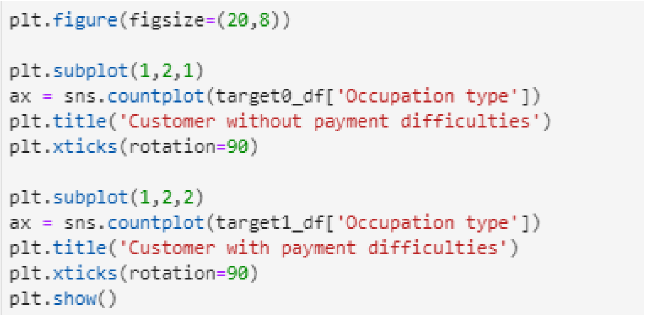


Figure 2.12 – Distribution of education types depending on availability of payment difficulties

Figure 2.13 shows that laborers are having more difficulties in repaying the loan and also the core staff and the sales staff. But in the case of laborers those who have without payment is way more then with having the payment.



```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
ax = sns.countplot(target0_df['Occupation type'])
plt.title('Customer without payment difficulties')
plt.xticks(rotation=90)

plt.subplot(1,2,2)
ax = sns.countplot(target1_df['Occupation type'])
plt.title('Customer with payment difficulties')
plt.xticks(rotation=90)
plt.show()
```

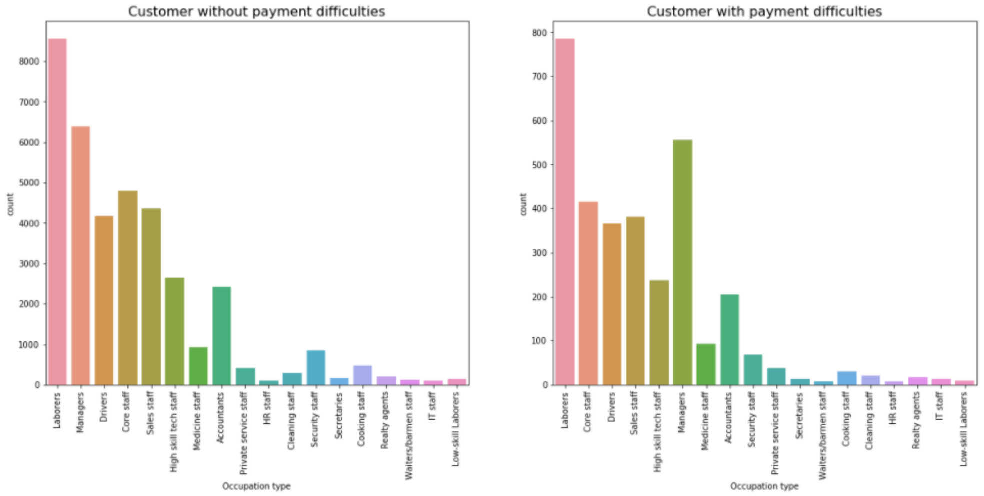


Figure 2.13 - Distribution of occupation types depending on availability of payment difficulties

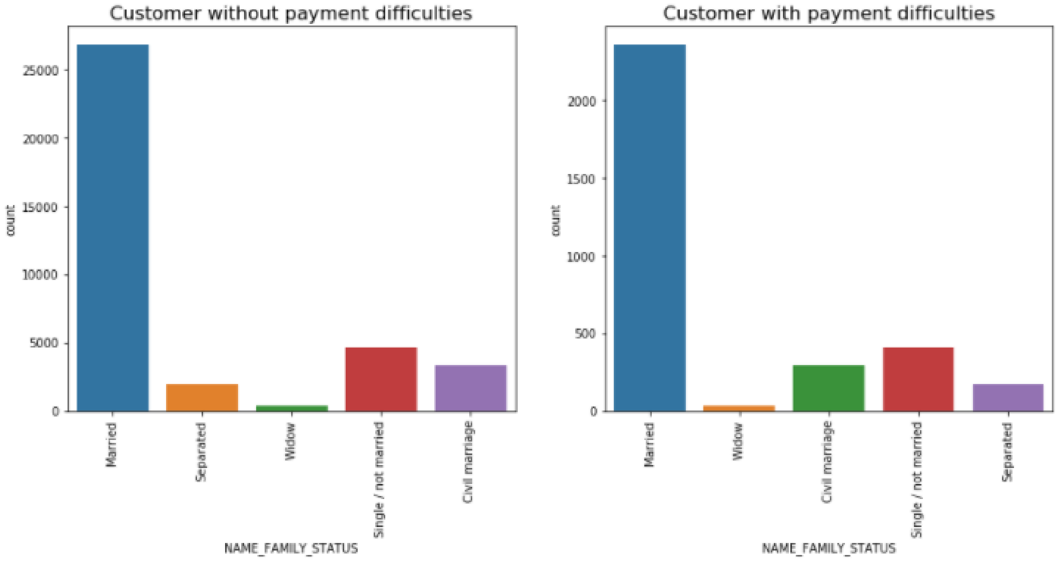


Figure 2.14 - Distribution of family status depending on availability of payment difficulties

Married applicants have the highest count with payment difficulties and both the FAMILY STATUS plots have a similar profile with respect to payment difficulties but differ in count (Figure 2.14).

Distribution of income type depending on availability of payment difficulties is described on the Figure 2.15.

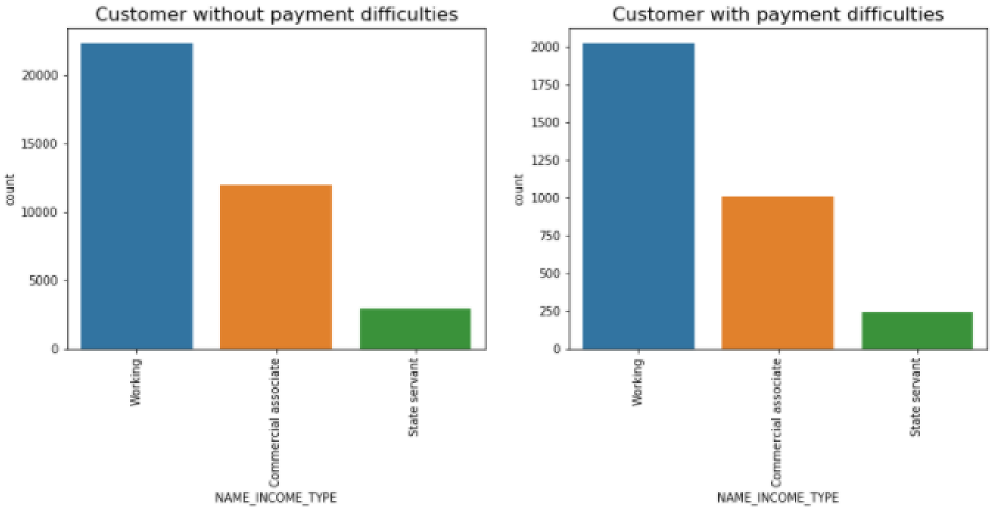


Figure 2.15 - Distribution of income type

Distribution of organization type depending on availability of payment difficulties is described on the Figure 2.16.

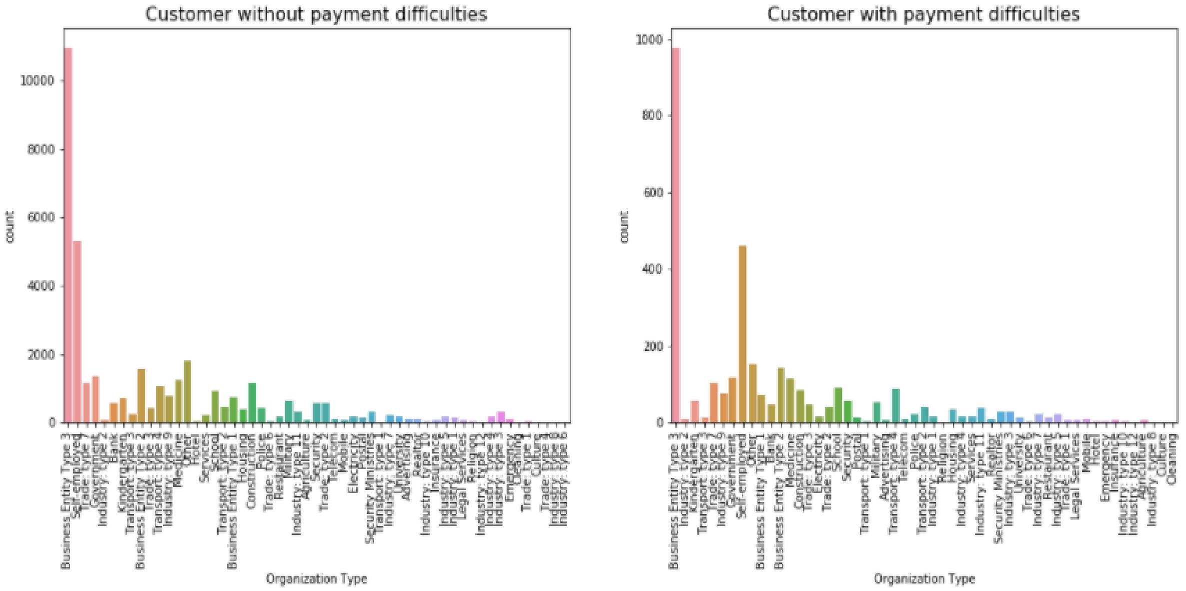
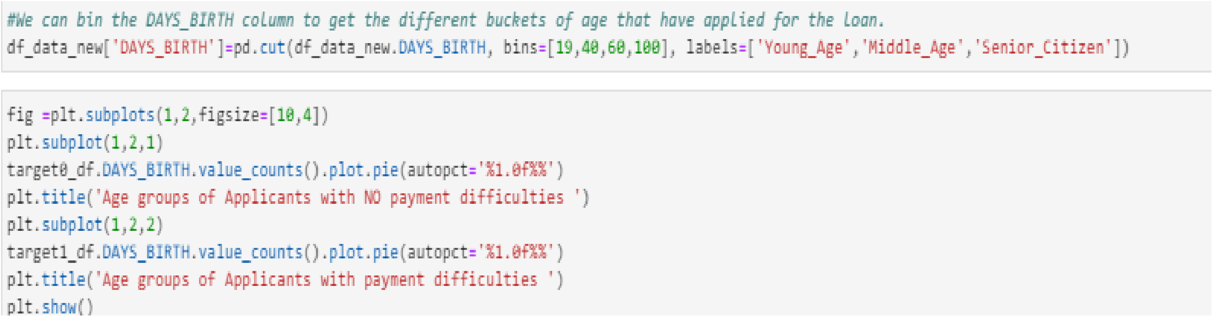


Figure 2.16 - Distribution of organization type

In order to consider the age range among banks client, the data need to be changed (Figure 2.17):



```
#We can bin the DAYS_BIRTH column to get the different buckets of age that have applied for the loan.
df_data_new['DAYS_BIRTH']=pd.cut(df_data_new.DAYS_BIRTH, bins=[19,40,60,100], labels=['Young_Age','Middle_Age','Senior_Citizen'])
```

```
fig =plt.subplots(1,2,figsize=[10,4])
plt.subplot(1,2,1)
target0_df.DAYS_BIRTH.value_counts().plot.pie(autopct='%1.0f%%')
plt.title('Age groups of Applicants with NO payment difficulties ')
plt.subplot(1,2,2)
target1_df.DAYS_BIRTH.value_counts().plot.pie(autopct='%1.0f%%')
plt.title('Age groups of Applicants with payment difficulties ')
plt.show()
```

Figure 2.17 – Changing age data

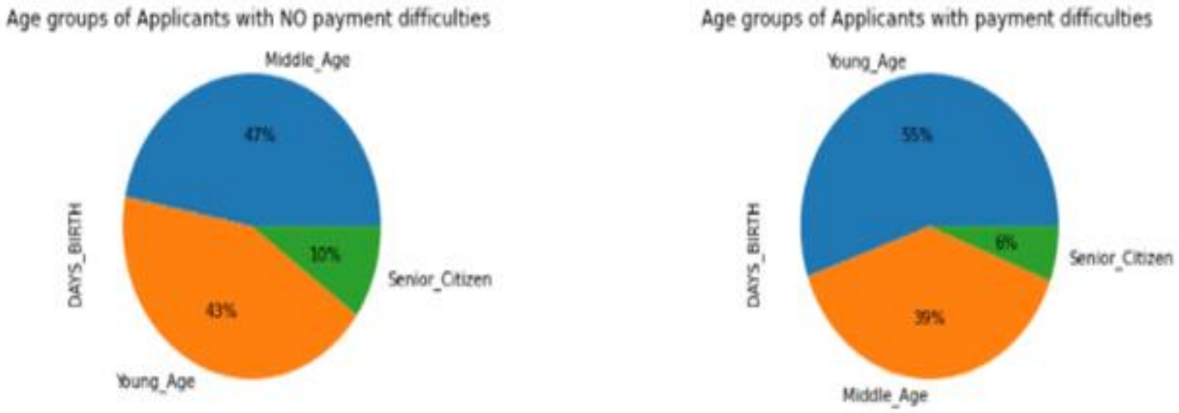


Figure 2.18 - Distribution of age groups depending on availability of payment difficulties

The Middle_Age group has the maximum percentage of customers WITHOUT payment difficulties, followed by the Young_Age and Senior_Citizen age groups. The Young_Age Group has the maximum percentage of customers with payment difficulties followed by the Middle_Age and Senior_Citizen age groups.

Comparing the two charts, we can conclude that the Middle_Age group has a lower payout risk compared to the Young_Age group (Figure. 2.18).

So Young_Age may have more payment issues.

People without payment difficuties take more credit for the annuity that they have (Figure 2.19):

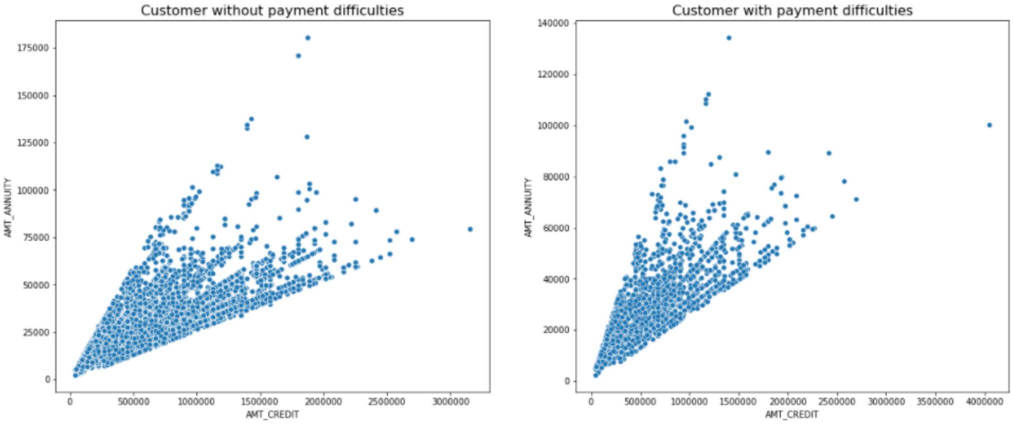
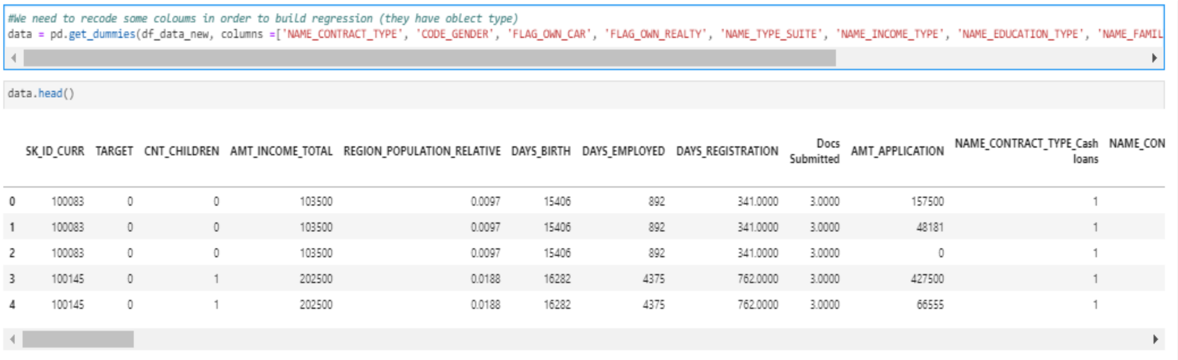


Figure 2.19 – Credit amount of the loan and loan annuity

In addition to numerical data, the database has categorical variables that cannot participate in the construction of mathematical models, so it is necessary to convert this data into numerical (Figure 2.20).



```
#We need to recode some coloums in order to build regression (they have oblect type)
data = pd.get_dummies(df_data_new, columns =['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMIL
```

```
data.head()
```

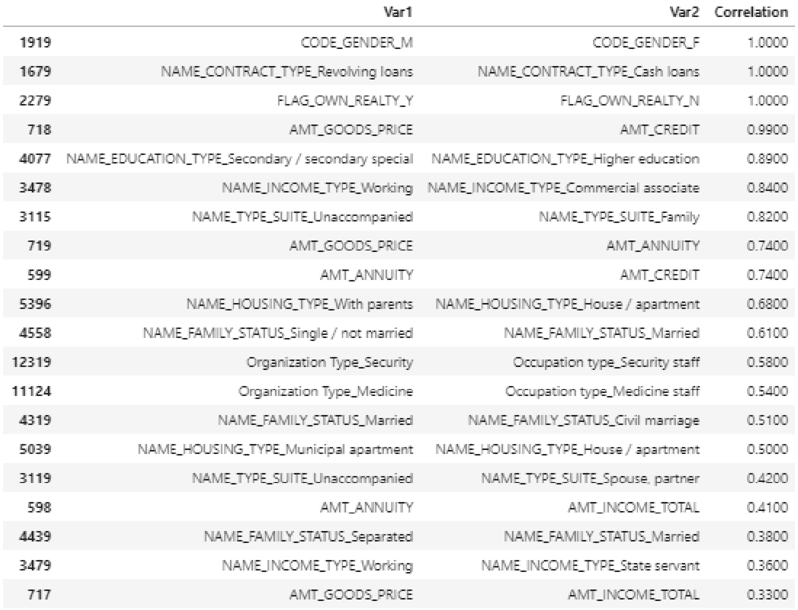
	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	Docs Submitted	AMT_APPLICATION	NAME_CONTRACT_TYPE_Cash loans	NAME_CON
0	100083	0	0	103500	0.0097	15406	892	341.0000	3.0000	157500	1	
1	100083	0	0	103500	0.0097	15406	892	341.0000	3.0000	48181	1	
2	100083	0	0	103500	0.0097	15406	892	341.0000	3.0000	0	1	
3	100145	0	1	202500	0.0188	16282	4375	762.0000	3.0000	427500	1	
4	100145	0	1	202500	0.0188	16282	4375	762.0000	3.0000	66555	1	

Figure 2.20 – Transformation object type data into float type

Now all the data is ready to participate in the construction of mathematical models to model the process of identifying bank customers for attempted credit fraud.

Correlation is a statistical quantity that measures the level of dependence between two random variables and is probabilistic [20].

To build mathematical models, you need to check the density of relationships between indicators. For adequate construction of models, the correlation coefficient should be less than 0.6.

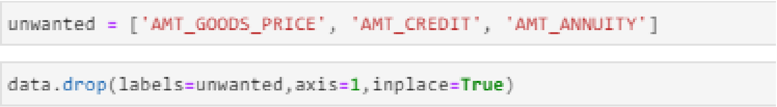


	Var1	Var2	Correlation
1919	CODE_GENDER_M	CODE_GENDER_F	1.0000
1679	NAME_CONTRACT_TYPE_Revolving loans	NAME_CONTRACT_TYPE_Cash loans	1.0000
2279	FLAG_OWN_REALTY_Y	FLAG_OWN_REALTY_N	1.0000
718	AMT_GOODS_PRICE	AMT_CREDIT	0.9900
4077	NAME_EDUCATION_TYPE_Secondary / secondary special	NAME_EDUCATION_TYPE_Higher education	0.8900
3478	NAME_INCOME_TYPE_Working	NAME_INCOME_TYPE_Commercial associate	0.8400
3115	NAME_TYPE_SUITE_Unaccompanied	NAME_TYPE_SUITE_Family	0.8200
719	AMT_GOODS_PRICE	AMT_ANNUITY	0.7400
599	AMT_ANNUITY	AMT_CREDIT	0.7400
5396	NAME_HOUSING_TYPE_With parents	NAME_HOUSING_TYPE_House / apartment	0.6800
4558	NAME_FAMILY_STATUS_Single / not married	NAME_FAMILY_STATUS_Married	0.6100
12319	Organization Type_Security	Occupation type_Security staff	0.5800
11124	Organization Type_Medicine	Occupation type_Medicine staff	0.5400
4319	NAME_FAMILY_STATUS_Married	NAME_FAMILY_STATUS_Civil marriage	0.5100
5039	NAME_HOUSING_TYPE_Municipal apartment	NAME_HOUSING_TYPE_House / apartment	0.5000
3119	NAME_TYPE_SUITE_Unaccompanied	NAME_TYPE_SUITE_Spouse, partner	0.4200
598	AMT_ANNUITY	AMT_INCOME_TOTAL	0.4100
4439	NAME_FAMILY_STATUS_Separated	NAME_FAMILY_STATUS_Married	0.3800
3479	NAME_INCOME_TYPE_Working	NAME_INCOME_TYPE_State servant	0.3600
717	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.3300

.

Figure 2.21 – Correlation between variables

The Figure 2.21-2.22 shows that variables such as AMT_GOODS_PRICE, AMT_CREDIT and AMT_ANNUNITY have close relationship between each other and correlations are more than 0.6, that`s why they must be removed from dataset for correct building mathematical models.



```
unwanted = ['AMT_GOODS_PRICE', 'AMT_CREDIT', 'AMT_ANNUITY']
```

```
data.drop(labels=unwanted,axis=1,inplace=True)
```

Figure 2.22 – Removing variable

The Table 2.1 shows follow variables which will be used for investigation [21]:

Table 2.1 - Description of input data

Name of variable	Description	Rule of variable	Type
TARGET	Target variable shows: 0 – client without payment difficulties; 1 – client with payment difficulties.	target	binary
CODE_GENDER	Gender of the client (male, female)	input	categorical
NAME_CONTRACT _TYPE	Identification if loan is cash or revolving	input	categorical
FLAG_OWN_CAR	Flag if the client owns a car	input	categorical
FLAG_OWN_REALTY	Flag if client owns a house or flat	input	categorical
CNT_CHILDREN	Number of children the client has	input	integer
AMT_INCOME_TOTAL	Income of the client	input	float
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan	input	categorical
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave)	input	categorical
NAME_EDUCATION_TYPE	Level of highest education the client achieved	input	categorical
NAME_FAMILY_STATUS	Family status of the client	input	categorical
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents etc)	input	categorical
REGION_POPULATION_RELATIVE	Normalized population of region where cl ient lives (higher number means the client lives in more populated regio n	input	float
Occupation type	What kind of occupation does the client have	input	categorical
DAYS_EMPLOYED	How many days before the application the person started current employment	inpur	integer

Now all the data is ready to participate in the construction of mathematical models to model the process of identifying bank customers for attempted credit fraud.

2.3 Mathematical models and their interpretation

Mathematical model is a system of mathematical relationships between indicators that describe the studied phenomena and processes in economic, social, political and other spheres of life [22].

Mathematical models are universal, they are a set of interconnected mathematical and logical expressions that reflect real processes. The results of such models need to be compared with the data of the real model. The purpose of comparing and comparing these models is to verify the adequacy of the data and further improve the model over time.

To model the process of fraud in the process of lending to bank customers will use the following models:

- logistic regression;
- decision tree;
- neural network;

Logistic regression (model) is a statistical regression method used when the dependent change is binary, ie can be 0 or 1. Logistic regression is a predictive analysis and is used to describe data and explain the relationship between one dependent factor (variable) and one or more independent [23].

Subsequently, the result of the probability of occurrence of a phenomenon is converted to binary value to make the prediction real, this result is assigned to the class to which it belongs, based on whether it is close or not to the class itself.

Let *x* be any continuous value whose domain is (-∞, ∞). If you plug *x* into the sigmoid function like (Formula 2.1):

$$\mathrm{P}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)} \tag{2.1}$$

The Figure 2.24 shows the form of the logistic regression:

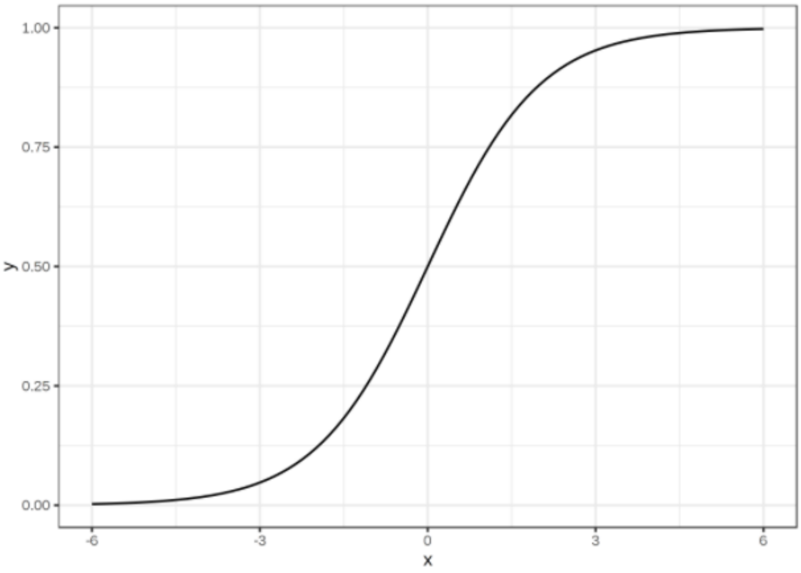


Figure 2.24 – Logistic model

The Formula 2.2 shows way from a linear regression to a logistic regression.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2.2}$$

For x like so (Formula 2.3):

$$\mathrm{P}(\hat{y} = 1) = \frac{1}{1+exp^{-\hat{y}}} = \frac{1}{1+exp^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{2.3}$$

This function reinterprets the OLS output as a probability. The formula above represents the output of a logistic regression model [24].

The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1. The weighted sum is transformed by the logistic function to a probability. Therefore equation needs to be reformulated for the interpretation so that only the linear term is on the right side of the formula 2.4.

$$\log\left(\frac{P(\hat{y}=1)}{1-P(\hat{y}=1)}\right) = \log\left(\frac{P(\hat{y}=1)}{P(\hat{y}=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2.4}$$

It is called the term in the log() function "odds" (probability of event divided by probability of no event) and wrapped in the logarithm it is called log odds.

This formula shows that the logistic regression model is a linear model for the log odds. To do this, the exp() should be applied to both sides of the equation. This is shown by formula 2.5.

$$\frac{P(\hat{y}=1)}{1-P(\hat{y}=1)} = odds = exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \qquad (2.5)$$

Then it need to be compared what happens when we increase one of the feature values by 1.

The ratio of the two predictions is described by formula 2.6.

$$\frac{odds_{x,+\,1}}{odds} = \frac{exp(\beta_0+\beta_1 x_1+\cdots+\beta_j\left(x_j+1\right)+\ldots+\beta_n x_n}{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_j x_j+\ldots+\beta_n x_n} \qquad (2.6)$$

To apply the following rule (Formula 2.7):

$$\frac{exp(a)}{exp(b)} = exp(a-b) \qquad (2.7)$$

And to remove many terms (Formula 2.8):

$$\frac{odds_{x,+1}}{odds} = exp\left(\beta_j\left(x_j+1\right) - \beta_1 x_j\right) = exp(\beta_j) \qquad (2.8)$$

In the end, the equation is something as simple as *exp()* of a feature weight. A change in a feature by one unit changes the odds ratio (multiplicative) by a factor of *exp(βj)*. It also could be interpreted by this way: A change in x_j by one unit increases the log odds ratio by the value of the corresponding weight [25].

The decision tree is one of the automatic methods of analysis and processing of huge data sets.

Using the decision tree method, you can combine it into three stages:

- data description - the use of this method allows you to store data in a convenient form and contains accurate descriptions of objects;

- classification - allows you to cope with the classification - the relationship of objects to each other;

- regression - allows you to determine the dependence of the target variable on independent (input) variables.

To make a decision or solve a certain problem, you need to:

- assess the state of the market, choosing specific factors X = (x1, x2, x3, .., xn). This step is performed by the user,

- determine the growth class of the decision tree from the upper levels to the lower. This stage is performed by the system [26].

First, the entire data set represented by the root vertex is taken to build the decision tree. Then the options for breaking down the data into branches corresponding to the root node are determined. These branches form a tree with the bark down. Methods of breaking down a set of data are called the decisive rule:

$$a_{ik} = \begin{cases} 1, if\ the\ condition\ is\ done; \\ \qquad 0,\ otherwise, \end{cases} \tag{2.9}$$

Where $a_{ik} = 1$, if condition s_i for rule r_k is done;

$S\{s_i\}, i = \overline{1, l}$– a set of conditions that describe the parameters of the selected subject area.

This rule is actually an "if,… then ..." algorithm and divides the set of records into two parts.

The neural network is an algorithm that combines biological principles and advanced statistics to solve such problems and problems in various fields. The neural network adopts a basic model of neural analogues interconnected in different ways.

Neural network is algorithms, which compute, from an input x, an output y (Figure 2.25).

The mathematical algorithm of the neural network is determined by the following function 2.10.

$$f_w(i.e. y = f_w(x)) \quad (2.10)$$

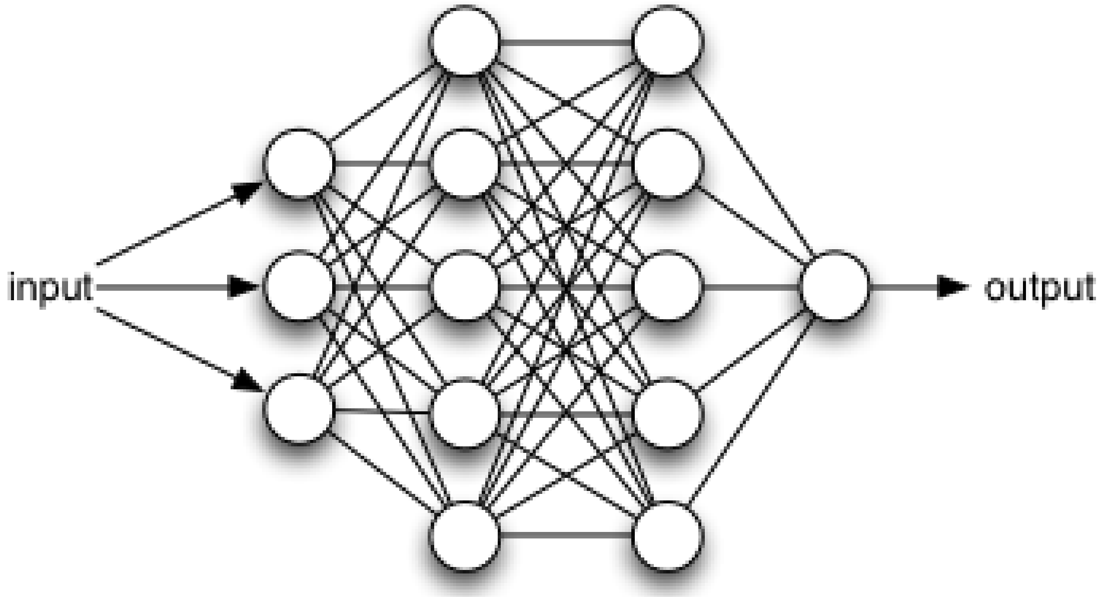


Figure 2.25 – Neural network

Taking into account the input and output variables presented on the Figure 2.25, the mathematical model of the neural network can be represented in general form as follows (Formulas 2.11-2.12):

$$h_1 = f(w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + \cdots + w_{18}^{(1)}x_{112} + b_1^{(1)} \quad (2.11)$$

$$h_2 = f(w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + \cdots + w_{28}^{(1)}x_{112} + b_2^{(1)} \quad (2.12)$$

By formulas 2.11-2.12 all hidden layers are calculated.

The computer program that calculates this function is very simple: it consists of a sequence of several stages, and each stage performs elementary calculations (addition, multiplication and maximum). But the big difference between the "classical" algorithm and the neural network is that the latter depends on the parameters that are the weight of the neurons.

This is done using mathematical and algorithmic methods called neural network "learning" and requires a lot of time, machine computing and energy [27]. The result of the comparison of the mathematical methods is described at the table 2.1.

Table 2.1 – The advantages and disadvantages of models

Name of the model	Advantages	Disadvantages
Logistic regression	one of the simplest algorithms; ease implementation and interpret; using predicted (train weight); ease to upgrade new data; is more efficient than linear regression.	can`t solve nonlinear problems; algorithm is sensitive to outliers; requires moderate or no multicollinearity between independent variables.
Decision Tree	requires less effort and time for data preparation; is very easy for explanation; doesn`t requires data normalization; allows missing data.	calculation can go far more complex than other algorithm; involves lots time for training the model; is inadequate for predicting continuous values.
Neural Network	is quite robust to noise in the training data; the errors in training set don`t effect on result; is used for fast evaluation of function.	requires parallel processing data; difficulties with showing the problem to the network; value can`t give optimal result.

All this methods are used for prediction some processes and phenomenon in the world. They are chosen depend on variables that dataset has and result.

CHAPTER 3 MODELING THE PROCESS OF IDENTIFYING BANK CUSTOMERS FOR CREDIT FRAUD

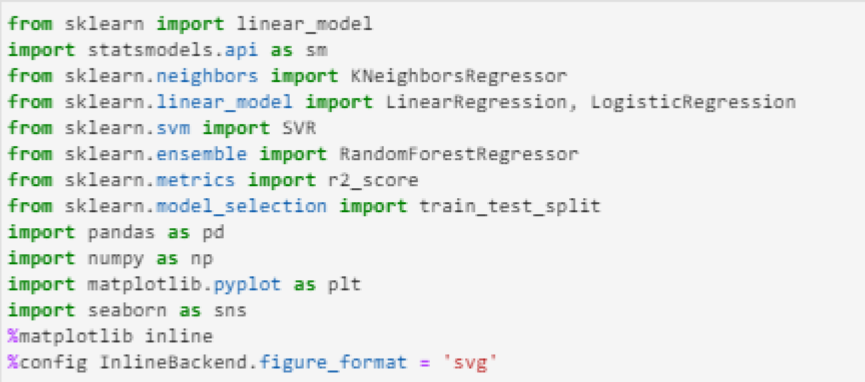
3.1 Construction of mathematical models for modeling the process of detecting credit fraud

3.1.1. Construction of the Logistic regression

The logistics model is built only under the conditions that the dependent variable is binary and with large sample sizes.

To construct the logistic regression, the data was prepared in chapter 2.2, the sample was cleared from unnecessary data, and the categorical metrics were converted to binary variables.

In order to start building a logistic regression, necessary libraries need to be imported (Figure. 3.1).

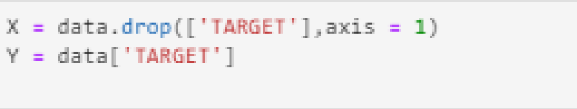


```
from sklearn import linear_model
import statsmodels.api as sm
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
%config InlineBackend.figure_format = 'svg'
```

Figure 3.1 – Libraries for building logistic regression

Now the investigation needs to define the dependent and independent variables. The dependent variable is TARGET, which shows the difficulty of repaying the loan, all other independent variables. These variables are presented in table 2.1.

The Figure 3.2 represents determination of dependent and independent variables.



```
X = data.drop(['TARGET'],axis = 1)
Y = data['TARGET']
```

Figure 3.2 –Determination of dependent and independent variables

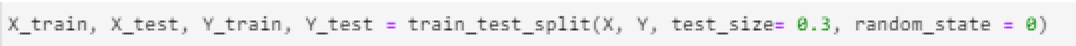
Before starting building a logistics model, database is needed to be divided into test and train data (Figure 3.3). These parts include:

– X_train - this set includes all independent variables that will be used to train the model. The amount of data usage for the test model is also specified (test_size = 0.3), which means that 70% of the data will be used for model training and 30% for model testing.

– X_test is the 30% data balance that will be used to test the data.

– Y_train is a dependent variable that should be provided by this model, it includes category labels against independent variables.

– Y_test - this data has category labels for test data , these labels will be used to check the accuracy between actual and predicted categories.



```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size= 0.3, random_state = 0)
```

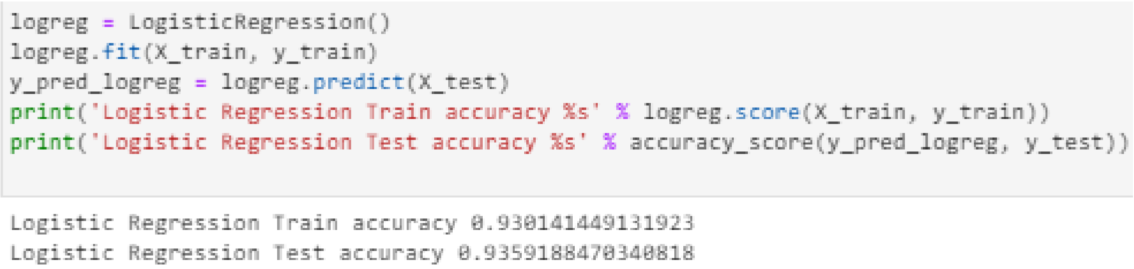
Figure 3.3 – Selection train and test splits

A logistic model can be built using the data that was distributed in the above operation.

First of all, it is necessary to calculate the forecast of the results of test and training sets.

Accuracy is the proportion of correct predictions over total predictions [28].

The results of the calculation of the accuracy for two sets are described on the figure 3.4.

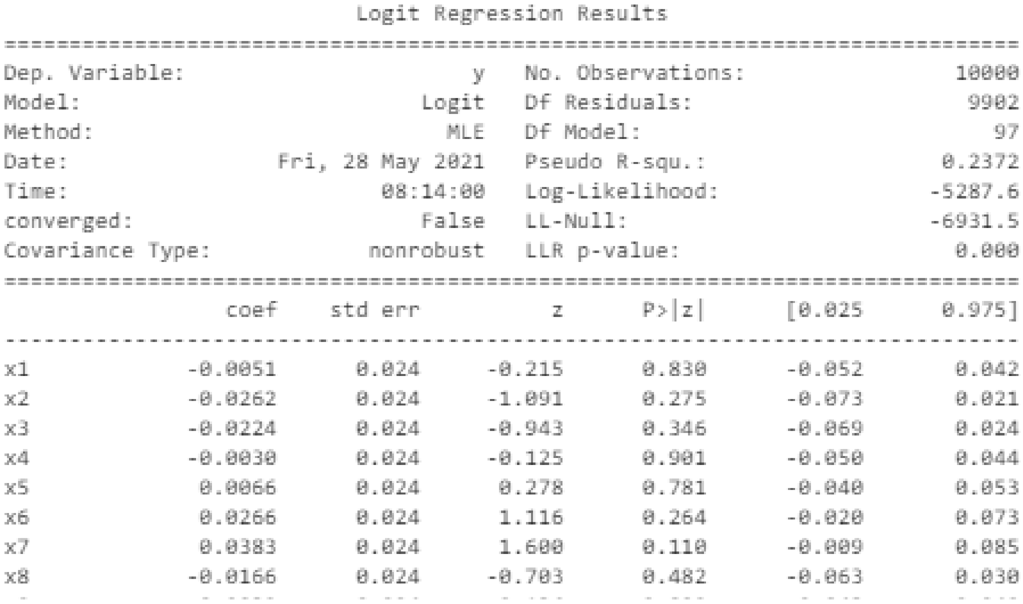


```
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred_logreg = logreg.predict(X_test)
print('Logistic Regression Train accuracy %s' % logreg.score(X_train, y_train))
print('Logistic Regression Test accuracy %s' % accuracy_score(y_pred_logreg, y_test))
```

```
Logistic Regression Train accuracy 0.930141449131923
Logistic Regression Test accuracy 0.9359188470340818
```

Figure 3.4 – Logistic regression train and test accuracy

The results show that the share of correct predictions in the training sample is 93% and in the test sample - 93.6%.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9902
Method:                           MLE   Df Model:                           97
Date:                Fri, 28 May 2021   Pseudo R-squ.:                  0.2372
Time:                        08:14:00   Log-Likelihood:                -5287.6
converged:                      False   LL-Null:                       -6931.5
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0051      0.024     -0.215      0.830      -0.052       0.042
x2            -0.0262      0.024     -1.091      0.275      -0.073       0.021
x3            -0.0224      0.024     -0.943      0.346      -0.069       0.024
x4            -0.0030      0.024     -0.125      0.901      -0.050       0.044
x5             0.0066      0.024      0.278      0.781      -0.040       0.053
x6             0.0266      0.024      1.116      0.264      -0.020       0.073
x7             0.0383      0.024      1.600      0.110      -0.009       0.085
x8            -0.0166      0.024     -0.703      0.482      -0.063       0.030
```

Figure 3.5 - Regression results

Figure 3.5 shows the results of the constructed logistic regression [35].

The coefficient of determination (R2) is a statistical indicator that represents the proportion of variance for the dependent variable, which is explained by a number of independent variables in the regression model [29].

According to the results of the regression logit, the coefficient of determination is equal to 0.2372, independent variables describe the target variable only by 23.72%. This means that there is a low level of correlation between the dependent and independent variable.

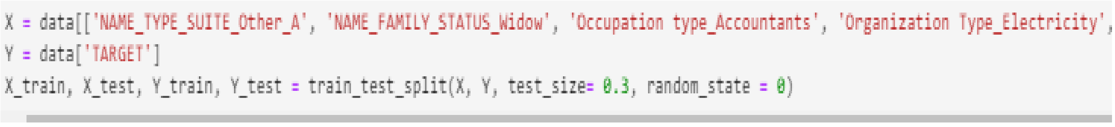
A p-value or probability value tells you how likely it is that your data could have arisen under the null hypothesis. This is done by calculating the probability of your test statistics, ie the amount calculated by the statistical test using your data.

P -values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the p -value, the greater the chance of rejecting the null hypothesis [30].

But looking at the p-value of each independent variable, we can conclude that almost none of indicators are not statistically significant (see Appendix B).

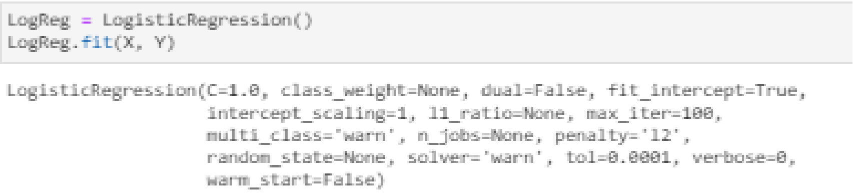
In order to create better model the significant variables need to be chosen. After investigation all variables by p-value, the logit regression has variables, such as (Figure 3.6-3.7):

– x_1 - 'NAME_TYPE_SUITE_Other_A';
– x_2 - 'NAME_FAMILY_STATUS_Widow';
– x_3 - 'Occupation type_Accountants';
– x_4 - 'Organization Type_Electricity';
– x_5 - 'Organization Type_Industry: type 3';
– x_6 - 'Organization Type_Military';
– x_7 - 'Organization Type_School'.



```
X = data[['NAME_TYPE_SUITE_Other_A', 'NAME_FAMILY_STATUS_Widow', 'Occupation type_Accountants', 'Organization Type_Electricity',
Y = data['TARGET']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size= 0.3, random_state = 0)
```

Figure 3.6 –Determination of dependent and independent variables

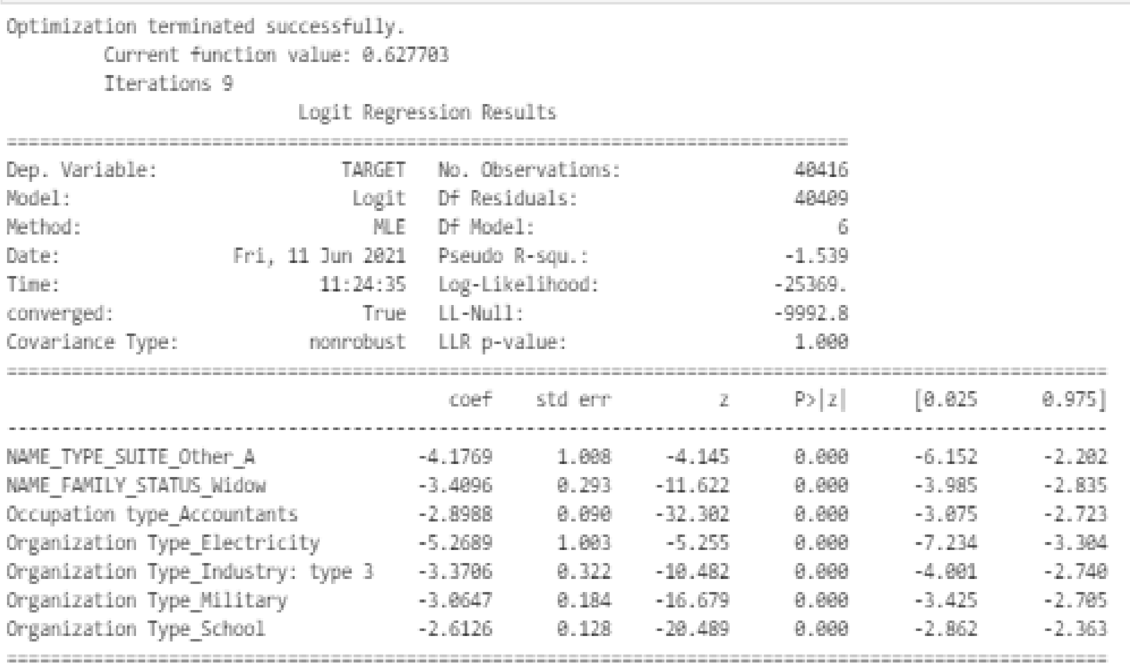


```
LogReg = LogisticRegression()
LogReg.fit(X, Y)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

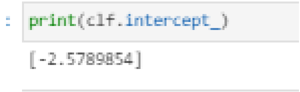
Figure 3.7 – Training of Logistic regression

The new logistical regression is created with listed independent variables (Figure 3.8):



```
Optimization terminated successfully.
         Current function value: 0.627703
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:                40416
Model:                          Logit   Df Residuals:                    40409
Method:                           MLE   Df Model:                            6
Date:                Fri, 11 Jun 2021   Pseudo R-squ.:                  -1.539
Time:                        11:24:35   Log-Likelihood:                -25369.
converged:                       True   LL-Null:                       -9992.8
Covariance Type:            nonrobust   LLR p-value:                     1.000
====================================================================================================
                                         coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
NAME_TYPE_SUITE_Other_A               -4.1769      1.008     -4.145      0.000      -6.152      -2.202
NAME_FAMILY_STATUS_Widow              -3.4096      0.293    -11.622      0.000      -3.985      -2.835
Occupation type_Accountants           -2.8988      0.090    -32.302      0.000      -3.075      -2.723
Organization Type_Electricity         -5.2689      1.003     -5.255      0.000      -7.234      -3.304
Organization Type_Industry: type 3    -3.3706      0.322    -10.482      0.000      -4.001      -2.740
Organization Type_Military            -3.0647      0.184    -16.679      0.000      -3.425      -2.705
Organization Type_School              -2.6126      0.128    -20.409      0.000      -2.862      -2.363
====================================================================================================
```

Figure 3.8 – Result of Logistic regression



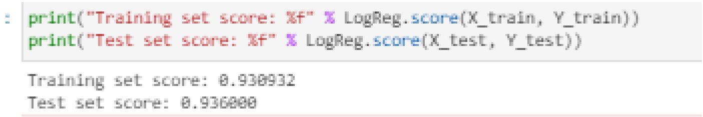
```
: print(clf.intercept_)

  [-2.5789854]
```

Figure 3.9 – Intercept coefficient for Logistic equation

The Figure 3.8 shows the new result of logistic regression. All variables are significant, because p-value is less than 0.05.

The figure 3.10 shows the results of calculation of accuracy. This model provides decent accuracy of 0.936 and 0.931 for the training and test data sets, respectively. There is a high accuracy of the model.

.



```
: print("Training set score: %f" % LogReg.score(X_train, Y_train))
  print("Test set score: %f" % LogReg.score(X_test, Y_test))

  Training set score: 0.930932
  Test set score: 0.936000
```

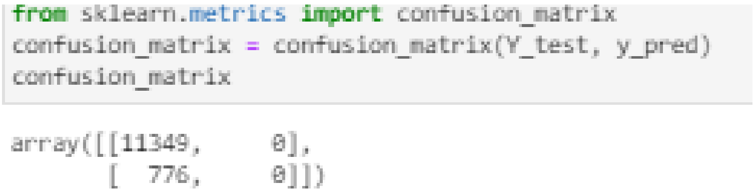
Figure 3.10 – Accuracy for train and test data

Analysis of the quality of the model, which predicts the probability of occurrence of a particular event, is determined primarily by how well it predicted the outcome. Such characteristics are determined quantitatively, in percentage, as well as by the coefficient of misclassification (*MisclassificationRate*) (Formula 3.1).

$$Misclassification\ Rate = \frac{\text{Number of incorrectly classified cases}}{\text{All cases}}. \qquad (3.1)$$

The indicator represented by formula 3.1 can be calculated using the Confusion matrix.

Confusion matrix is a table used to assess the effectiveness of model classification, ie the basis for constructing this matrix is to calculate the number of correct and incorrect forecasts [31]. The result of calculation of the confusion matrix is presented on the figure 3.11.



```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(Y_test, y_pred)
confusion_matrix
```

```
array([[11349,     0],
       [  776,     0]])
```

Figure 3.11 - Confusion matrix for Logistic regression

Figure 3.11 depicts the Confusion matrix as an array. The size of this matrix is 2x2, because the model is a binary classification (0 and 1).

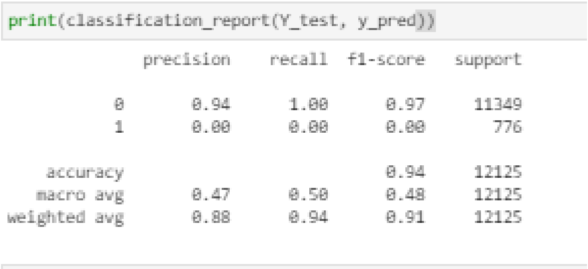
Diagonal values represent accurate predictions, while non-diagonal elements represent inaccurate predictions. That is, the constructed model has 11349 correct predictions and 776 incorrect ones.

The classification report displays the precision, recall, F1, and support scores for the model (Figure 3.12).

Precision is the ratio of true positives to the sum of true positives and false positives. In this case, 94% of the bank's customers are not fraudsters.

Recall is the proportion of positive cases that have been correctly identified.

The F1-score is a weighted harmonic mean of accuracy and response.



```
print(classification_report(Y_test, y_pred))

              precision    recall  f1-score   support

           0       0.94      1.00      0.97     11349
           1       0.00      0.00      0.00       776

    accuracy                           0.94     12125
   macro avg       0.47      0.50      0.48     12125
weighted avg       0.88      0.94      0.91     12125
```

Figure 3.12 – Classification report

The receiver operating characteristic curve (ROC) is another popular tool used with binary classifiers.

An ROC curve is a graph showing the performance of a classification model at all classification thresholds [32].

The result of building of ROC curve is presented on the figure 3.13.

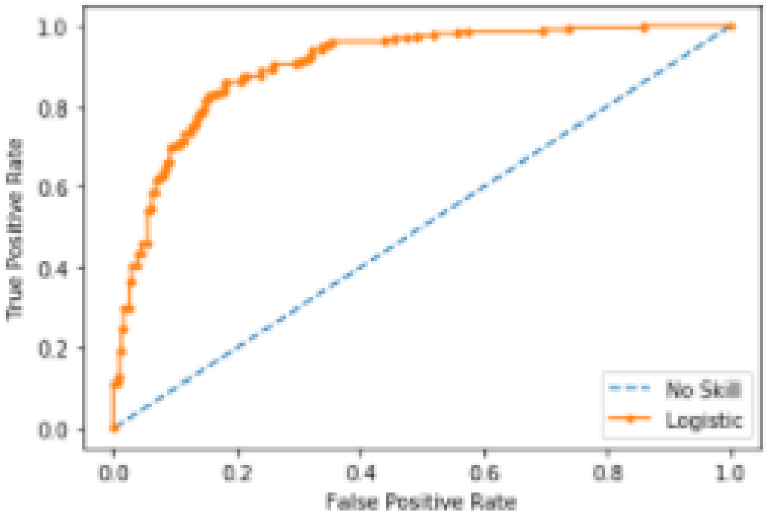


Figure 3.13 - ROC curve for logistic regression

The Figure 3.13 shows that logistic model has a true positive rate, because good classifier remains as far away from it as possible (towards the upper left corner) [36].

Despite the high level of p-value of each of the independent variables, which shows the insignificance of the variables, ROC curve shows a high dependence of the number of correctly classified positive examples on the number of incorrectly classified negative examples.

Mathematically, the model for detecting the probability of fraudulent operation can be represented by formula 3.2:

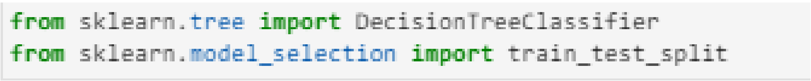
$$P = \frac{1}{1+E^{-2.579-4.1769x_1-3.4096x_2-2.8988x_3-5.2689x_4-3.3706x_5-3.0647x_6-2.6126x_7}} \qquad (3.2)$$

Predictive estimates of Exp of this model indicate that when you change the value of the independent variables x_1-x_7 by 1, the probability of fraudulent operation is not traced.

The conducted investigation and modeling logistic regression don`t give excellent result and not indicate influence of independent variables on target variable. Logistic regression is not suitable for given dataset. So, this model is not recommended to use in this case.

3.1.2. Construction of the Decision Tree

Building a decision tree also begins with importing the necessary libraries (Figure 3.14) [37]:

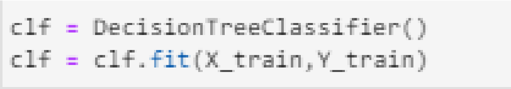


```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
```

Figure 3.14 – Importing libraries

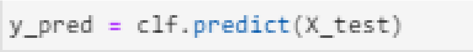
The next step is to define the dependent and independent variables, as well as the division into test and training parts of the database. These steps were performed in the implementation of logistic regression (Figures 3.2 - 3.3).

To build a decision tree, the Scikit-Learn library is used and to teach the .fit model using training data (Figures 3.15-3.16):



```
clf = DecisionTreeClassifier()
clf = clf.fit(X_train,Y_train)
```

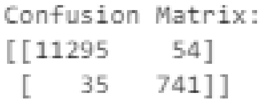
Figure 3.15 – Creating and model training



```
y_pred = clf.predict(X_test)
```

Figure 3.16 – Determination of prediction of target variable

Figure 3.17 depicts the Confusion matrix, which shows that 11295 and 741 records are correct predictions and 35 and 54 are incorrect.



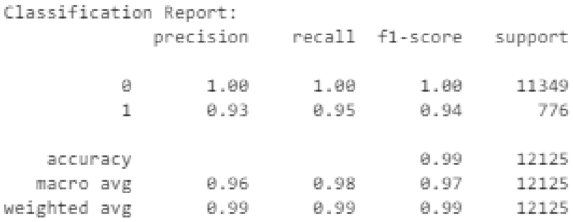
```
Confusion Matrix:
[[11295    54]
 [   35   741]]
```

Figure 3.17 – Confusion matrix for Decision Tree

Figure 3.17 depicts the Classification report for a better understanding of the relationship between the dependent variable and the independent ones. (client with difficulty in paying the loan) is 93%.

In an imbalanced classification problem with two classes, recall is calculated as the number of true positives divided by the total number of true positives and false negatives. In this case the result shows that recalls are perfect for both option of target variable (close to 1).

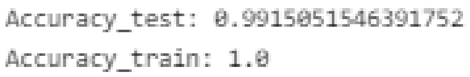
The same situation is with f1-score. The result shows that the model is balanced (Figure 3.18).



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     11349
           1       0.93      0.95      0.94       776

    accuracy                           0.99     12125
   macro avg       0.96      0.98      0.97     12125
weighted avg       0.99      0.99      0.99     12125
```

Figure 3.18 – Classification report for Decision tree

The Figure 3.19 shows the result of calculating of accuracy of test dataset and train that are equal to 0.9915 and 1.0, It means that precision of the forecast almost is 100%.



```
Accuracy_test: 0.9915051546391752
Accuracy_train: 1.0
```

Figure 3.19– Accuracy for Decision tree

The result of the building of the Decision Tree is given on the figure 3.20 [Appendix D].

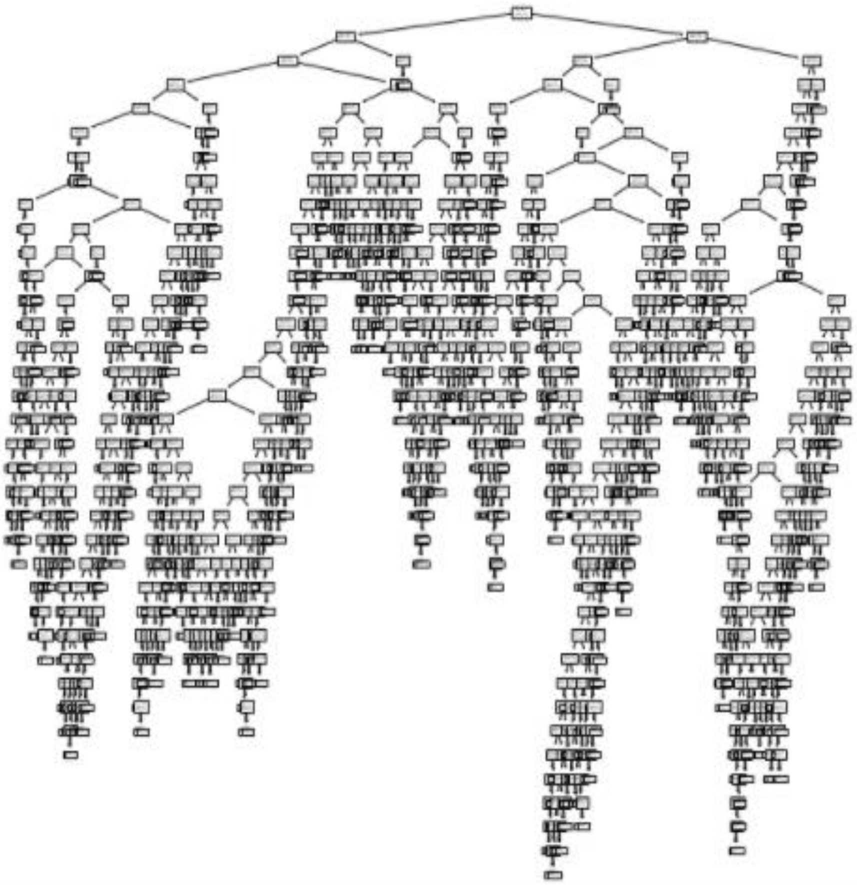
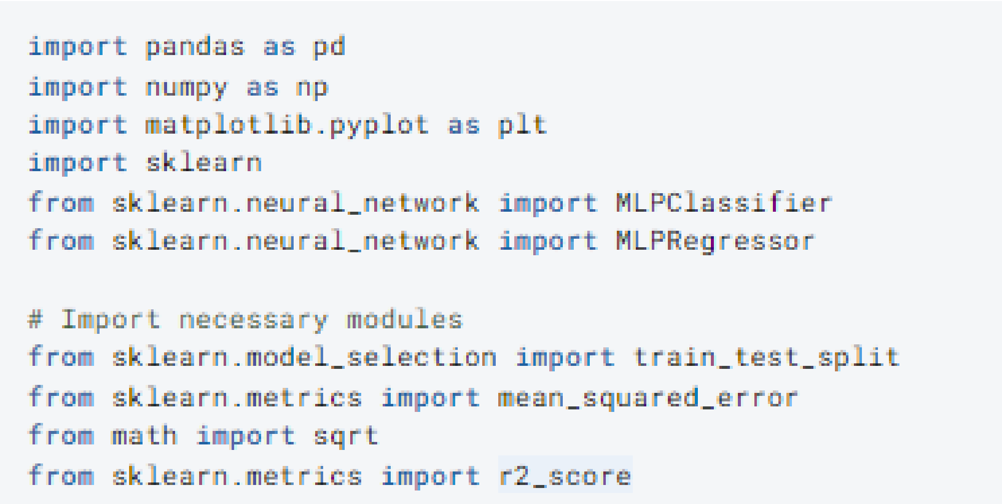


Figure 3.20 – Decision Tree

Having constructed a decision tree it is possible to draw a conclusion that the decision tree has high accuracy of forecasting of the target variable ‘TARGET’ [38].

3.1.3 Construction of the Neural Network

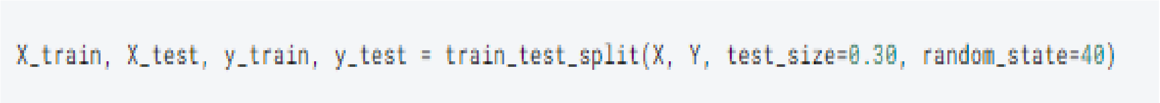
Building a neural network also begins with importing the necessary libraries (Figure. 3.21) [39]:



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn
from sklearn.neural_network import MLPClassifier
from sklearn.neural_network import MLPRegressor

# Import necessary modules
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt
from sklearn.metrics import r2_score
```

Figure 3.21– Importing necessary libraries

The next step is to create test and training data, dividing the database into two parts at a percentage of 70% and 30%, respectively (Figure 3.22).

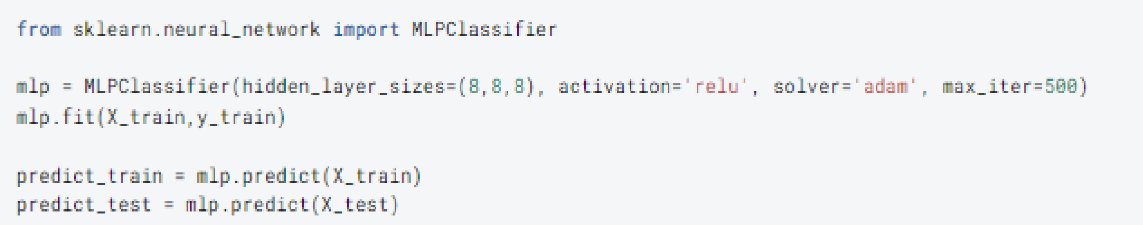


```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=40)
```

Figure 3.22 – Creating the Training and Test Datasets

At this stage, the scikit-learn library's estimator object ‘Multi-Layer Perceptron Classifier’ is used for building neural network.

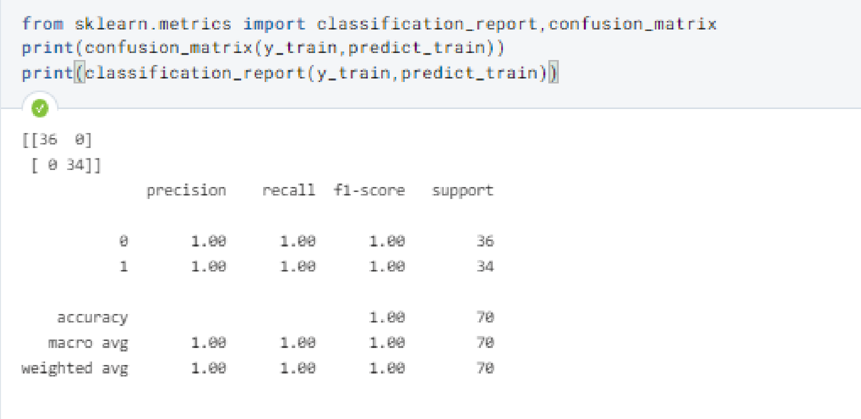
The model is also trained and the trained model is used to create predictions on training and test data (Figure 3.23).



```
from sklearn.neural_network import MLPClassifier

mlp = MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu', solver='adam', max_iter=500)
mlp.fit(X_train,y_train)

predict_train = mlp.predict(X_train)
predict_test = mlp.predict(X_test)
```

Figure 3.23 – Evaluating the Neural Network Model

The following result, which is given on the figure 3.24, shows the effectiveness of the model on training data. This means that the training data perfectly describe this constructed model.

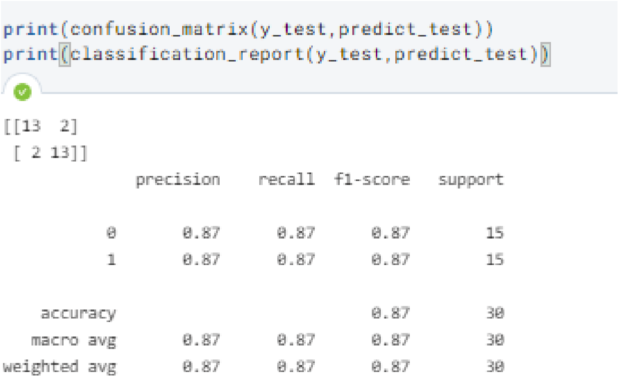


```
from sklearn.metrics import classification_report,confusion_matrix
print(confusion_matrix(y_train,predict_train))
print(classification_report(y_train,predict_train))
```

```
[[36  0]
 [ 0 34]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        36
           1       1.00      1.00      1.00        34

    accuracy                           1.00        70
   macro avg       1.00      1.00      1.00        70
weighted avg       1.00      1.00      1.00        70
```

Figure 3.24 – Classification report and Confusion matrix for training dataset

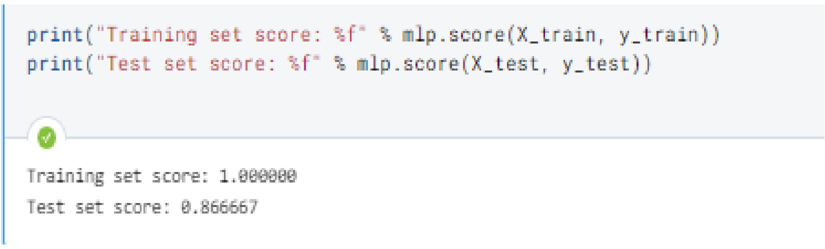
The following result shows, which is given on the figure 3.25, the effectiveness of the model on test data. Accuracy and evaluation are 87%, which shows an almost perfect result of describing the data with a test data set.



```
print(confusion_matrix(y_test,predict_test))
print(classification_report(y_test,predict_test))
```

```
[[13  2]
 [ 2 13]]
              precision    recall  f1-score   support

           0       0.87      0.87      0.87        15
           1       0.87      0.87      0.87        15

    accuracy                           0.87        30
   macro avg       0.87      0.87      0.87        30
weighted avg       0.87      0.87      0.87        30
```

Figure 3.25 – Classification report and Confusion matrix for testing dataset



```
print("Training set score: %f" % mlp.score(X_train, y_train))
print("Test set score: %f" % mlp.score(X_test, y_test))
```

```
Training set score: 1.000000
Test set score: 0.866667
```

Figure 3.26 – Calculation of accuracy for test and train dataset

The figure 3.26 shows the results of calculation of accuracy. This model provides decent accuracy of 100% and 86.7% for the training and test data sets, respectively. There is a high accuracy of the model, which is higher than the base 66%. This model can be improved by changing the arguments in the neural network evaluator or by cross-checking.

The figure 3.27 shows the result of building ROC curve for Neural Network.

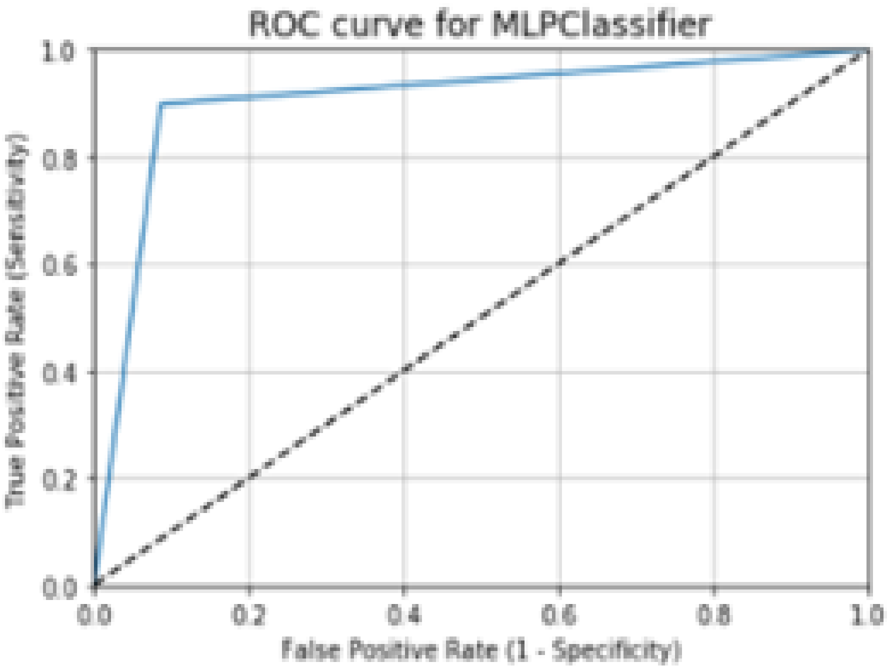
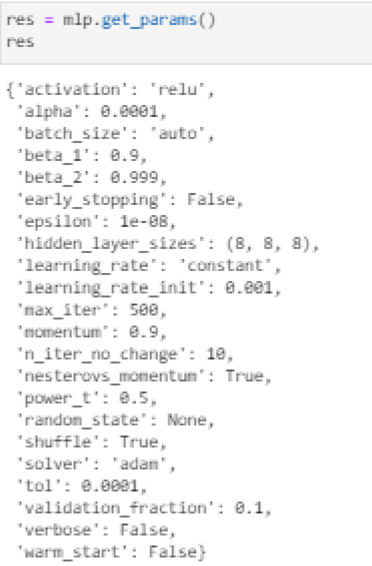


Figure 3.27 – ROC curve for Neural Network

Figure 3.27 shows Schematic representation of a neural network.

As a result, a neural network consisting of eight hidden layers was generated (Figure 2.28):



```
res = mlp.get_params()
res
```

```
{'activation': 'relu',
 'alpha': 0.0001,
 'batch_size': 'auto',
 'beta_1': 0.9,
 'beta_2': 0.999,
 'early_stopping': False,
 'epsilon': 1e-08,
 'hidden_layer_sizes': (8, 8, 8),
 'learning_rate': 'constant',
 'learning_rate_init': 0.001,
 'max_iter': 500,
 'momentum': 0.9,
 'n_iter_no_change': 10,
 'nesterovs_momentum': True,
 'power_t': 0.5,
 'random_state': None,
 'shuffle': True,
 'solver': 'adam',
 'tol': 0.0001,
 'validation_fraction': 0.1,
 'verbose': False,
 'warm_start': False}
```

Figure 2.28 – Parameters of neural network

Figure 2.29 shows the Schematic representation of a neural network.

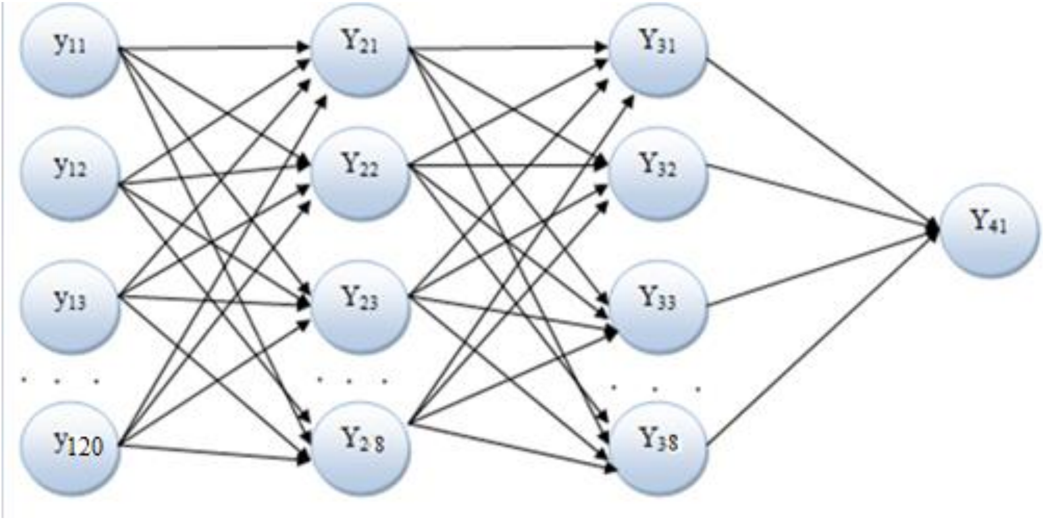
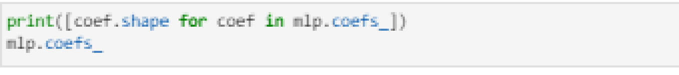


Figure 3.29 – Schematic representation of a neural network

In order to present mathematical form of the neural network, the weight coefficient of the hidden layers need to be found with help follow operation (Figure 3.29):



```
print([coef.shape for coef in mlp.coefs_])
mlp.coefs_
```

Figure 3.29 – Finding neural network weight coefficient

All found coefficient are shown in Appendix C.

Mathematically, the model for detecting the probability of fraudulent operation can be represented by formula 3.3 – 3.10:

$$y_{11} = -0.1161x_1 - 0.5501x_2 + 0.3569x_3 + \cdots + 0.2065x_8 \quad (3.3)$$

$$y_{12} = -0.6962x_1 + 0.3309x_2 + 0.2792x_3 + \cdots + 0.2065x_8 \quad (3.4)$$

$$y_{120} = 0.3499x_1 - 0.4575x_2 + 0.0761x_3 + \cdots + 0.3483x_8 \quad (3.5)$$

$$y_{21} = -0.3926x_1 - 0.6920x_2 + \cdots - 0.1789x_8 \quad (3.6)$$

$$y_{28} = 0.5144x_1 - 0.0822x_2 + \cdots - 0.2014x_8 \quad (3.7)$$

$$y_{31} = -4.4977x_1 + 9.19e^{-01}x_2 + 0.27 + \cdots - 3.5448e^{-01}x_8 \quad (3.8)$$

$$y_{38} = 4.44e^{-02}x_1 - 3.4965e^{-02} + \cdots + 3.4229e^{01}x_8 \quad (3.9)$$

$$y_{41} = 0.2948x_1 - 0.92x_2 + \cdots - 0.6388x_8 \quad (3.10)$$

3.2 Analysis of the quality and accuracy of the constructed models and their comparison

The precision of the model, in addition to accuracy, is the standard error.

The standard error (MSE) is the standard deviation that measures the difference between the predicted value and the actual.

The standard error is measured by formula (3.11) [34]:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}, \quad (3.11)$$

where $\hat{y}_i$ – simulated value (predicted probability of occurrence of the event of interest to the researcher);

y_i – the actual value of the indicator;

n – serial numder [33].

After building the models, Logistic regression, decision tree and neural network can be compared with each other for the accuracy and MSE of the description of the models training and test data set [40].

Table 3.1 lists the mathematical models in order from best to worst in quantification of accuracy and standard deviation. Therefore, the models describe the training models better than the test set.

Table 3.1 — Comparative characteristics of the coefficients of quality and accuracy of models: regression, decision tree and neural network

Name of model	Accuracy, %		MSE	
	Test dataset	Train dataset	Test dataset	Train dataset
Decision Tree	99,15	100	0.008	0.0
Neural Networks	86,7	100	0.089	0.004
Logistic regression	96,6	93	0.064	0.069

With this table we can conclude that the mathematical model - the decision tree best describes the likelihood of fraud in the process of lending to bank customers. The accuracy of the mathematical model of the decision tree is close to 1 and the standard deviation is close to zero. In second place was the Neural Network and in third place Logistic Regression. The latest models also have a high level of accuracy and the standard deviation differs only in hundredths. Modeling logistic regression don`t give excellent result and not indicate influence of independent variables on target variable. Logistic regression is not suitable for given dataset. So, this model is not recommended to use in this case.

CONCLUSION

In this worl was investigated and analyzed the probability of fraud in the process of lending to bank customers. The essence of credit fraud, credit risk and methods of its management were revealed.

Analysis and modeling were performed to assess the probability of credit fraud by building mathematical models, namely Logistics Regression, Decision Tree and Neural Network. We believe that the built model of the Decision Tree will be one of the best methods to determine the probability of fraud in the process of lending to bank customers with high accuracy.

Fifteen input variables were selected and a mathematical interpretation of logistic regression, decision tree, and neural network was presented to build a model for estimating the probability of credit fraud. The result was the construction of these three models using the Python programming language. The constructed models were tested for adequacy and quality, and their results were analyzed.

It turned out that the best in terms of adequacy and accuracy is the Decision Tree. According to this model, all indicators describe the model almost 100%.

The construction of the model is high quality and accurate for early assessment of the probability of fraud in the process of lending to bank customers.

As a result of using this model in financial institutions and banks, positive economic and social effects will be obtained.

All the tasks set to reveal the essence of the object of study - fraudulent transactions related to lending to bank customers, analysis of existing approaches to modeling and assessment of credit risks and building mathematical models, have been achieved.

REFERENCES

1. Dexterity: top schemes of fraud in banks [Electronic resource]. URL: https://home.kpmg/ua/uk/blogs/home/posts/2019/10/sprytnist-ruk-shahraystva-v-bankah.html.

2. How the Banking Sector Impacts Our Economy [Electronic resource]. URL: https://www.investopedia.com/ask/answers/032315/what-banking-sector.asp#:~:text=The%20banking%20sector%20is%20an,way%20to%20create%20more%20wealth.

3. Classification of financial fraud in a commercial bank [Electronic resource]. URL: http://www.ej.kherson.ua/journal/economic_23/3/23.pdf.

4. Credit Card Dump [Electronic resource]. URL: https://www.investopedia.com/terms/c/credit-card-dump.asp.

5. Top Five Types of Identity Theft, 2020 [Electronic resource]. URL: https://www.iii.org/table-archive/20279.

6. Credit fraud in the field of financial and credit relations [Electronic resource]. URL: http://ndekc.te.ua/news/kreditn-mahnats-v-sfer-fnansovokreditnih-vdnosin-ta-zahist-nteresv-h-subktv.

7. What is consumer credit? [Electronic resource]. URL: https://www.bankrate.com/glossary/c/consumer-credit/.

8. What Is Credit Fraud? [Electronic resource]. URL: https://www.experian.com/blogs/ask-experian/credit-education/preventing-fraud/what-is-credit-fraud/.

9. What is Credit Risk?[Electronic resource]. URL: https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-risk.

10. Credit risk management principles, tools and techniques [Electronic resource]. URL: https://www.theglobaltreasurer.com/2019/02/07/credit-risk-management-principles-tools-and-techniques/

11. Credit risk management methods [Electronic resource]. URL: https://buklib.net/books/32558/.

12. An Introduction to Credit Risk[Electronic resource]. URL: https://www.bookstime.com/articles/credit-risk.

13. Decree on approval of the Regulations on the organization of the risk management system in banks of Ukraine and banking groups [Electronic resource]. URL: https://zakon.rada.gov.ua/laws/show/v0064500-18#Text.

14. SCORING AS AN EXPERT EVALUATION METHOD CREDIT RISK OF A COMMERCIAL BANK IN CONSUMER LENDING [Electronic resource]. URL: https://web.znu.edu.ua/herald/issues/2008/econom_2008_1/2008-26-06/volik.pdf.

15. How Bank Scoring Works [Electronic resource]. URL: https://finance.ua/credits/kak-rabotaet-bankovskiy-skoring.

16. Scoring technologies for assessing the risks of the shahraya in the bank's activity [Electronic resource]. URL: https://onlinebank.dp.ua/publications/508-skoringovi-tekhnologiji-otsinyuvannya-rizikiv-shakhrajstva-v-bankivskij-diyalnosti/.

17. How bank scoring works [Electronic resource]. URL: https://finance.ua/credits/kak-rabotaet-bankovskiy-skoring.

18. Methods credit fraud detection [Electronic resource]. URL: http://dspace.wunu.edu.ua/bitstream/.PDF.

19. Credit Card Fraud Detection [Electronic resource]. URL: https://www.kaggle.com/mishra5001/credit-card?select=application_data.csv.

20. What is EDA? [Electronic resource]. URL: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm.

21. Credit Card Fraud Detection [Electronic resource]. URL: https://www.kaggle.com/mishra5001/credit-card?select=columns_description.csv.

22. Mathematical Models [Electronic resource]. URL: https://www.sciencedirect.com/topics/engineering/mathematical-model.

23. What is Logistic Regression? [Electronic resource]. URL: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.

24. Logistic regression theory for practitioners [Electronic resource]. URL: https://towardsdatascience.com/the-data-scientists-field-guide-to-logistic-regression-part-1-intuition-97084b11bd68.

25. Logistic regression [Electronic resource]. URL: https://christophm.github.io/interpretable-ml-book/logistic.html.

26. Using the decision tree [Electronic resource]. URL: https://pidru4niki.com/10780621/ekonomika/vikoristannya_dereva_rishen.

27. Mathematics of Neural Networks [Electronic resource]. URL: https://mathematical-tours.github.io/book-basics-sources/neural-networks-en/NeuralNetworksEN.pdf.

28. Logistic Regression, Accuracy, and Cross-Validation [Electronic resource]. URL: https://medium.com/@lily_su/logistic-regression-accuracy-cross-validation-58d9eb58d6e6.

29. R-squared definition [Electronic resource]. URL: https://www.investopedia.com/terms/r/r-squared.asp.

30. P-value explained [Electronic resource]. URL: https://www.scribbr.com/statistics/p-value.

31. Understanding Logistic Regression in Python [Electronic resource]. URL: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

32. Classification: ROC Curve and AUC [Electronic resource]. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

33. Quantitative criteria for assessing the accuracy of a number of equivalent measurements of one quantity [Electronic resource]. URL: http://kaf-gis.kh.ua/33-kilkisni-kriteriyi-ocinyuvannya-tochnosti-ryadu-rivnotochnih-vimiriv-odniieyi-velichini.

34. Mean Squared Error: Definition and Example [Electronic resource]. URL: https://www.statisticshowto.com/probability-and-statistics/statistics definitions/mean-squared-error/.

35. Logistic Regression in Python [Electronic resource]. URL:https://realpython.com/logistic-regression-python/.

36. ROC Curve and AUC From Scratch in NumPy (Visualized!) [Electronic resource]. URL:https://realpython.com/logistic-regression-python/https://towardsdatascience.com/roc-curve-and-auc-from-scratch-in-numpy-visualized-2612bb9459ab.

37. Decision Tree Classification in Python [Electronic resource]. URL: https://www.datacamp.com/community/tutorials/decision-tree-classification-python.

38. Visualize a Decision Tree in 4 Ways with Scikit-Learn and Python [Electronic resource]. URL: https://mljar.com/blog/visualize-decision-tree/.

39. How to create simple Neural Network in Python [Electronic resource]. URL: https://www.kdnuggets.com/2018/10/simple-neural-network-python.html

40. MSE in Python[Electronic resource]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

APPENDIXES

Appendix A
(mandatory)
ABSTRACT OF QUALIFICATION WORK

SUMMARY

Radko VV Assessment of the probability of fraud in the process of lending to bank customers. Qualification work of the bachelor: special. 051 Economics (Business Analytics) / Supervisor H.M. Yarovenko. Sumy: Sumy State University, 2021.

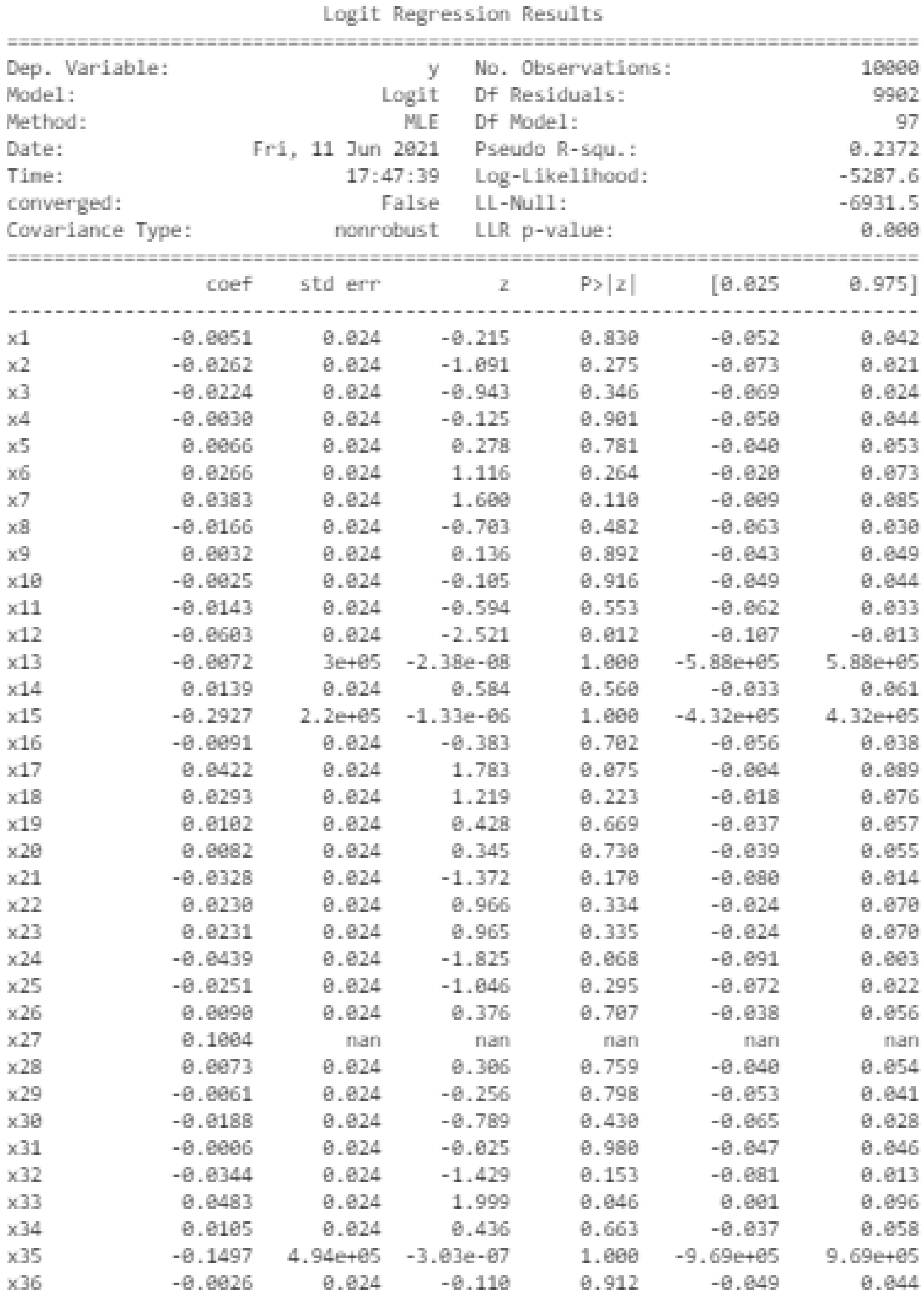
The process of modeling the probability of fraud in the process of lending to bank customers is investigated. An analysis of the current state of credit fraud was conducted. Modern approaches to combating credit risks were analyzed. A conceptual model for estimating the probability of fraud has been built. Mathematical models were built, such as logistic regression, decision tree and neural network. As a result, the Decision Tree was identified as the best method that can effectively and accurately prevent the likelihood of fraud in the process of lending to bank customers.

Keywords: credit fraud, credit risk, scoring, Python, logistic regression, decision tree, neural network, ROC-curve.

Appendix B

(informative)

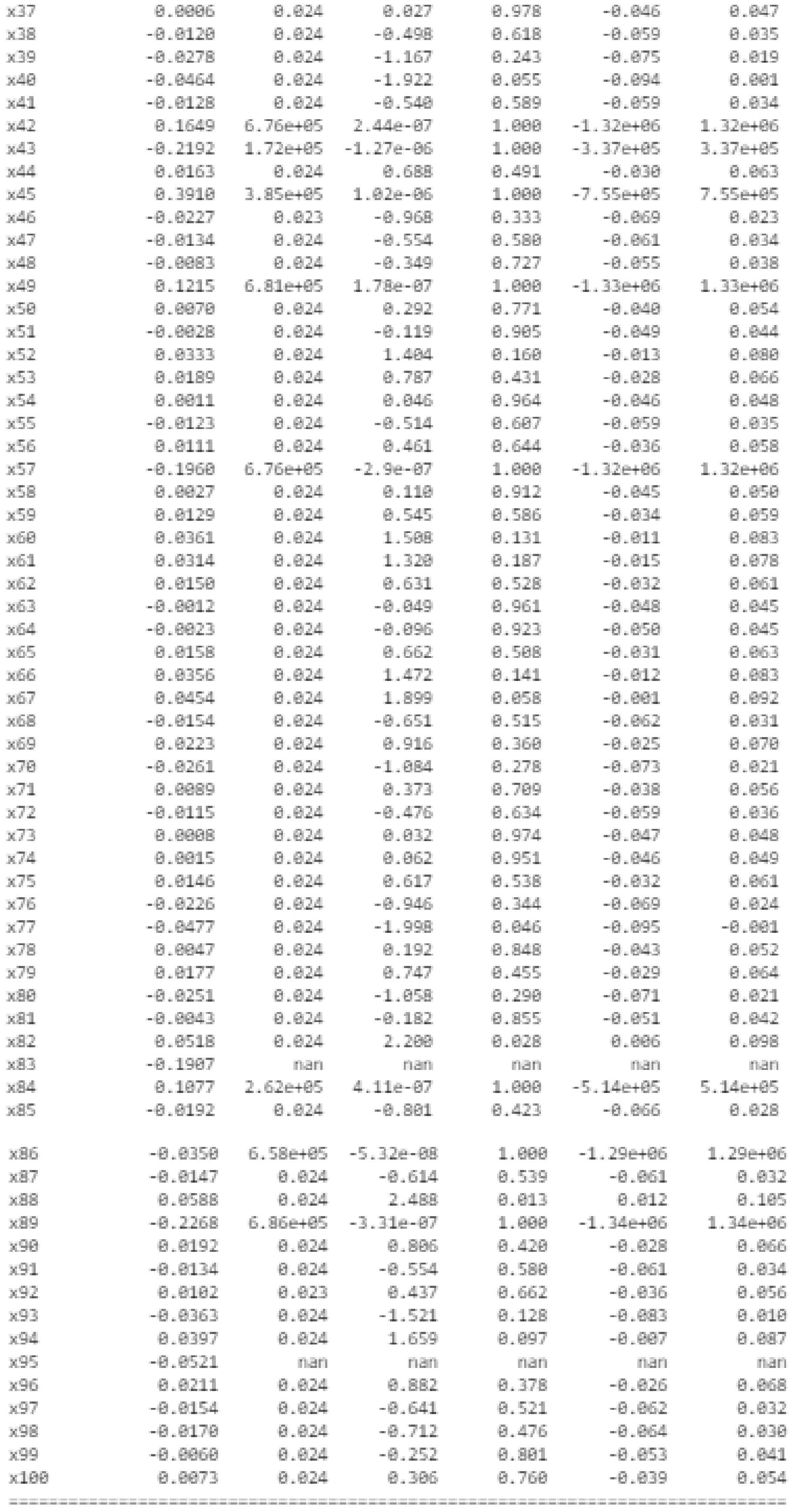
PRIMARY RESULT OF LOGISTIC REGRESSION



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9902
Method:                           MLE   Df Model:                           97
Date:                Fri, 11 Jun 2021   Pseudo R-squ.:                  0.2372
Time:                        17:47:39   Log-Likelihood:                -5287.6
converged:                      False   LL-Null:                       -6931.5
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0051      0.024     -0.215      0.830      -0.052       0.042
x2            -0.0262      0.024     -1.091      0.275      -0.073       0.021
x3            -0.0224      0.024     -0.943      0.346      -0.069       0.024
x4            -0.0030      0.024     -0.125      0.901      -0.050       0.044
x5             0.0066      0.024      0.278      0.781      -0.040       0.053
x6             0.0266      0.024      1.116      0.264      -0.020       0.073
x7             0.0383      0.024      1.600      0.110      -0.009       0.085
x8            -0.0166      0.024     -0.703      0.482      -0.063       0.030
x9             0.0032      0.024      0.136      0.892      -0.043       0.049
x10           -0.0025      0.024     -0.105      0.916      -0.049       0.044
x11           -0.0143      0.024     -0.594      0.553      -0.062       0.033
x12           -0.0603      0.024     -2.521      0.012      -0.107      -0.013
x13           -0.0072      3e+05  -2.38e-08      1.000   -5.88e+05    5.88e+05
x14            0.0139      0.024      0.584      0.560      -0.033       0.061
x15           -0.2927    2.2e+05  -1.33e-06      1.000   -4.32e+05    4.32e+05
x16           -0.0091      0.024     -0.383      0.702      -0.056       0.038
x17            0.0422      0.024      1.783      0.075      -0.004       0.089
x18            0.0293      0.024      1.219      0.223      -0.018       0.076
x19            0.0102      0.024      0.428      0.669      -0.037       0.057
x20            0.0082      0.024      0.345      0.730      -0.039       0.055
x21           -0.0328      0.024     -1.372      0.170      -0.080       0.014
x22            0.0230      0.024      0.966      0.334      -0.024       0.070
x23            0.0231      0.024      0.965      0.335      -0.024       0.070
x24           -0.0439      0.024     -1.825      0.068      -0.091       0.003
x25           -0.0251      0.024     -1.046      0.295      -0.072       0.022
x26            0.0090      0.024      0.376      0.707      -0.038       0.056
x27            0.1004        nan        nan        nan         nan         nan
x28            0.0073      0.024      0.306      0.759      -0.040       0.054
x29           -0.0061      0.024     -0.256      0.798      -0.053       0.041
x30           -0.0188      0.024     -0.789      0.430      -0.065       0.028
x31           -0.0006      0.024     -0.025      0.980      -0.047       0.046
x32           -0.0344      0.024     -1.429      0.153      -0.081       0.013
x33            0.0483      0.024      1.999      0.046       0.001       0.096
x34            0.0105      0.024      0.436      0.663      -0.037       0.058
x35           -0.1497   4.94e+05  -3.03e-07      1.000   -9.69e+05    9.69e+05
x36           -0.0026      0.024     -0.110      0.912      -0.049       0.044
```

Figure B.1 – Result of Logistic regression

Continuation of appendix B

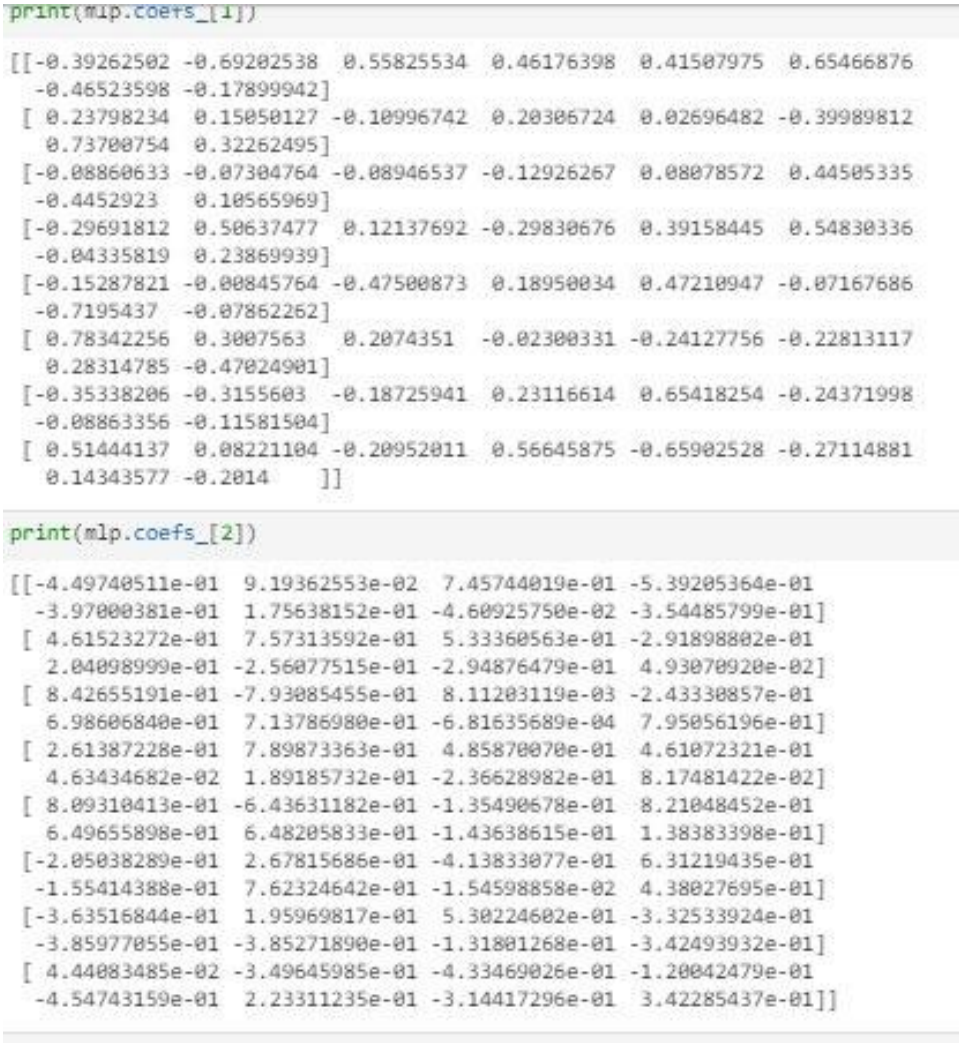


```
x37           0.0006      0.024      0.027      0.978      -0.046       0.047
x38          -0.0120      0.024     -0.498      0.618      -0.059       0.035
x39          -0.0278      0.024     -1.167      0.243      -0.075       0.019
x40          -0.0464      0.024     -1.922      0.055      -0.094       0.001
x41          -0.0128      0.024     -0.540      0.589      -0.059       0.034
x42           0.1649   6.76e+05   2.44e-07      1.000   -1.32e+06    1.32e+06
x43          -0.2192   1.72e+05  -1.27e-06      1.000   -3.37e+05    3.37e+05
x44           0.0163      0.024      0.688      0.491      -0.030       0.063
x45           0.3910   3.85e+05   1.02e-06      1.000   -7.55e+05    7.55e+05
x46          -0.0227      0.023     -0.968      0.333      -0.069       0.023
x47          -0.0134      0.024     -0.554      0.580      -0.061       0.034
x48          -0.0083      0.024     -0.349      0.727      -0.055       0.038
x49           0.1215   6.81e+05   1.78e-07      1.000   -1.33e+06    1.33e+06
x50           0.0070      0.024      0.292      0.771      -0.040       0.054
x51          -0.0028      0.024     -0.119      0.905      -0.049       0.044
x52           0.0333      0.024      1.404      0.160      -0.013       0.080
x53           0.0189      0.024      0.787      0.431      -0.028       0.066
x54           0.0011      0.024      0.046      0.964      -0.046       0.048
x55          -0.0123      0.024     -0.514      0.607      -0.059       0.035
x56           0.0111      0.024      0.461      0.644      -0.036       0.058
x57          -0.1960   6.76e+05   -2.9e-07      1.000   -1.32e+06    1.32e+06
x58           0.0027      0.024      0.110      0.912      -0.045       0.050
x59           0.0129      0.024      0.545      0.586      -0.034       0.059
x60           0.0361      0.024      1.508      0.131      -0.011       0.083
x61           0.0314      0.024      1.320      0.187      -0.015       0.078
x62           0.0150      0.024      0.631      0.528      -0.032       0.061
x63          -0.0012      0.024     -0.049      0.961      -0.048       0.045
x64          -0.0023      0.024     -0.096      0.923      -0.050       0.045
x65           0.0158      0.024      0.662      0.508      -0.031       0.063
x66           0.0356      0.024      1.472      0.141      -0.012       0.083
x67           0.0454      0.024      1.899      0.058      -0.001       0.092
x68          -0.0154      0.024     -0.651      0.515      -0.062       0.031
x69           0.0223      0.024      0.916      0.360      -0.025       0.070
x70          -0.0261      0.024     -1.084      0.278      -0.073       0.021
x71           0.0089      0.024      0.373      0.709      -0.038       0.056
x72          -0.0115      0.024     -0.476      0.634      -0.059       0.036
x73           0.0008      0.024      0.032      0.974      -0.047       0.048
x74           0.0015      0.024      0.062      0.951      -0.046       0.049
x75           0.0146      0.024      0.617      0.538      -0.032       0.061
x76          -0.0226      0.024     -0.946      0.344      -0.069       0.024
x77          -0.0477      0.024     -1.998      0.046      -0.095      -0.001
x78           0.0047      0.024      0.192      0.848      -0.043       0.052
x79           0.0177      0.024      0.747      0.455      -0.029       0.064
x80          -0.0251      0.024     -1.058      0.290      -0.071       0.021
x81          -0.0043      0.024     -0.182      0.855      -0.051       0.042
x82           0.0518      0.024      2.200      0.028       0.006       0.098
x83          -0.1907        nan        nan        nan         nan         nan
x84           0.1077   2.62e+05   4.11e-07      1.000   -5.14e+05    5.14e+05
x85          -0.0192      0.024     -0.801      0.423      -0.066       0.028

x86          -0.0350   6.58e+05  -5.32e-08      1.000   -1.29e+06    1.29e+06
x87          -0.0147      0.024     -0.614      0.539      -0.061       0.032
x88           0.0588      0.024      2.488      0.013       0.012       0.105
x89          -0.2268   6.86e+05  -3.31e-07      1.000   -1.34e+06    1.34e+06
x90           0.0192      0.024      0.806      0.420      -0.028       0.066
x91          -0.0134      0.024     -0.554      0.580      -0.061       0.034
x92           0.0102      0.023      0.437      0.662      -0.036       0.056
x93          -0.0363      0.024     -1.521      0.128      -0.083       0.010
x94           0.0397      0.024      1.659      0.097      -0.007       0.087
x95          -0.0521        nan        nan        nan         nan         nan
x96           0.0211      0.024      0.882      0.378      -0.026       0.068
x97          -0.0154      0.024     -0.641      0.521      -0.062       0.032
x98          -0.0170      0.024     -0.712      0.476      -0.064       0.030
x99          -0.0060      0.024     -0.252      0.801      -0.053       0.041
x100          0.0073      0.024      0.306      0.760      -0.039       0.054
==============================================================================
```

Figure B.2 – Result of Logistic regression

Appendix C

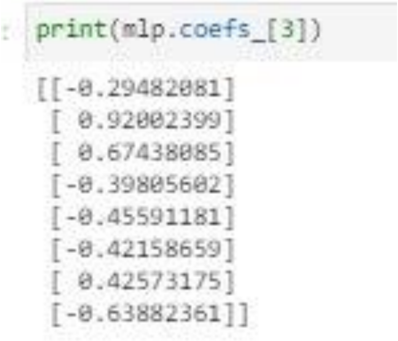
WEIGHT COEFFICIENT FOR NEURAL NETWORK



```
print(mlp.coefs_[1])

[[-0.39262502 -0.69202538  0.55825534  0.46176398  0.41507975  0.65466876
  -0.46523598 -0.17899942]
 [ 0.23798234  0.15050127 -0.10996742  0.20306724  0.02696482 -0.39989812
   0.73700754  0.32262495]
 [-0.08860633 -0.07304764 -0.08946537 -0.12926267  0.08078572  0.44505335
  -0.4452923   0.10565969]
 [-0.29691812  0.50637477  0.12137692 -0.29830676  0.39158445  0.54830336
  -0.04335819  0.23869939]
 [-0.15287821 -0.00845764 -0.47500873  0.18950034  0.47210947 -0.07167686
  -0.7195437  -0.07862262]
 [ 0.78342256  0.3007563   0.2074351  -0.02300331 -0.24127756 -0.22813117
   0.28314785 -0.47024901]
 [-0.35338206 -0.3155603  -0.18725941  0.23116614  0.65418254 -0.24371998
  -0.08863356 -0.11581504]
 [ 0.51444137  0.08221104 -0.20952011  0.56645875 -0.65902528 -0.27114881
   0.14343577 -0.2014    ]]
```

```
print(mlp.coefs_[2])

[[-4.49740511e-01  9.19362553e-02  7.45744019e-01 -5.39205364e-01
  -3.97000381e-01  1.75638152e-01 -4.60925750e-02 -3.54485799e-01]
 [ 4.61523272e-01  7.57313592e-01  5.33360563e-01 -2.91898802e-01
   2.04098999e-01 -2.56077515e-01 -2.94876479e-01  4.93070920e-02]
 [ 8.42655191e-01 -7.93085455e-01  8.11203119e-03 -2.43330857e-01
   6.98606840e-01  7.13786980e-01 -6.81635689e-04  7.95056196e-01]
 [ 2.61387228e-01  7.89873363e-01  4.85870070e-01  4.61072321e-01
   4.63434682e-02  1.89185732e-01 -2.36628982e-01  8.17481422e-02]
 [ 8.09310413e-01 -6.43631182e-01 -1.35490678e-01  8.21048452e-01
   6.49655898e-01  6.48205833e-01 -1.43638615e-01  1.38383398e-01]
 [-2.05038289e-01  2.67815686e-01 -4.13833077e-01  6.31219435e-01
  -1.55414388e-01  7.62324642e-01 -1.54598858e-02  4.38027695e-01]
 [-3.63516844e-01  1.95969817e-01  5.30224602e-01 -3.32533924e-01
  -3.85977055e-01 -3.85271890e-01 -1.31801268e-01 -3.42493932e-01]
 [ 4.44083485e-02 -3.49645985e-01 -4.33469026e-01 -1.20042479e-01
  -4.54743159e-01  2.23311235e-01 -3.14417296e-01  3.42285437e-01]]
```

Figure C.1 – Weight coefficient for Neural network



```
print(mlp.coefs_[3])

[[-0.29482081]
 [ 0.92002399]
 [ 0.67438085]
 [-0.39805602]
 [-0.45591181]
 [-0.42158659]
 [ 0.42573175]
 [-0.63882361]]
```

Figure C.2 – Weight coefficient for Neural network

Appendix D

(informative)

DECISION TREE

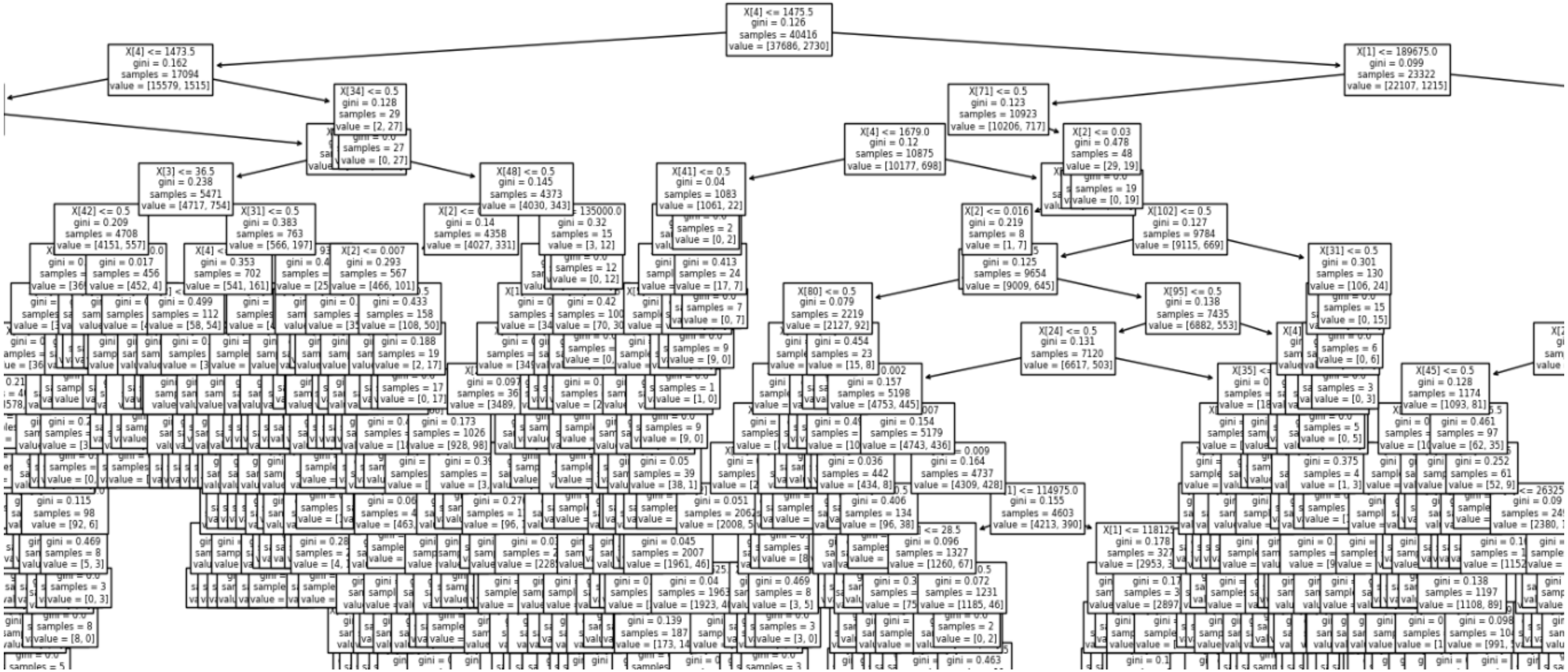


Figure D.1 – First part of Decision tree

Continuation of appendix D

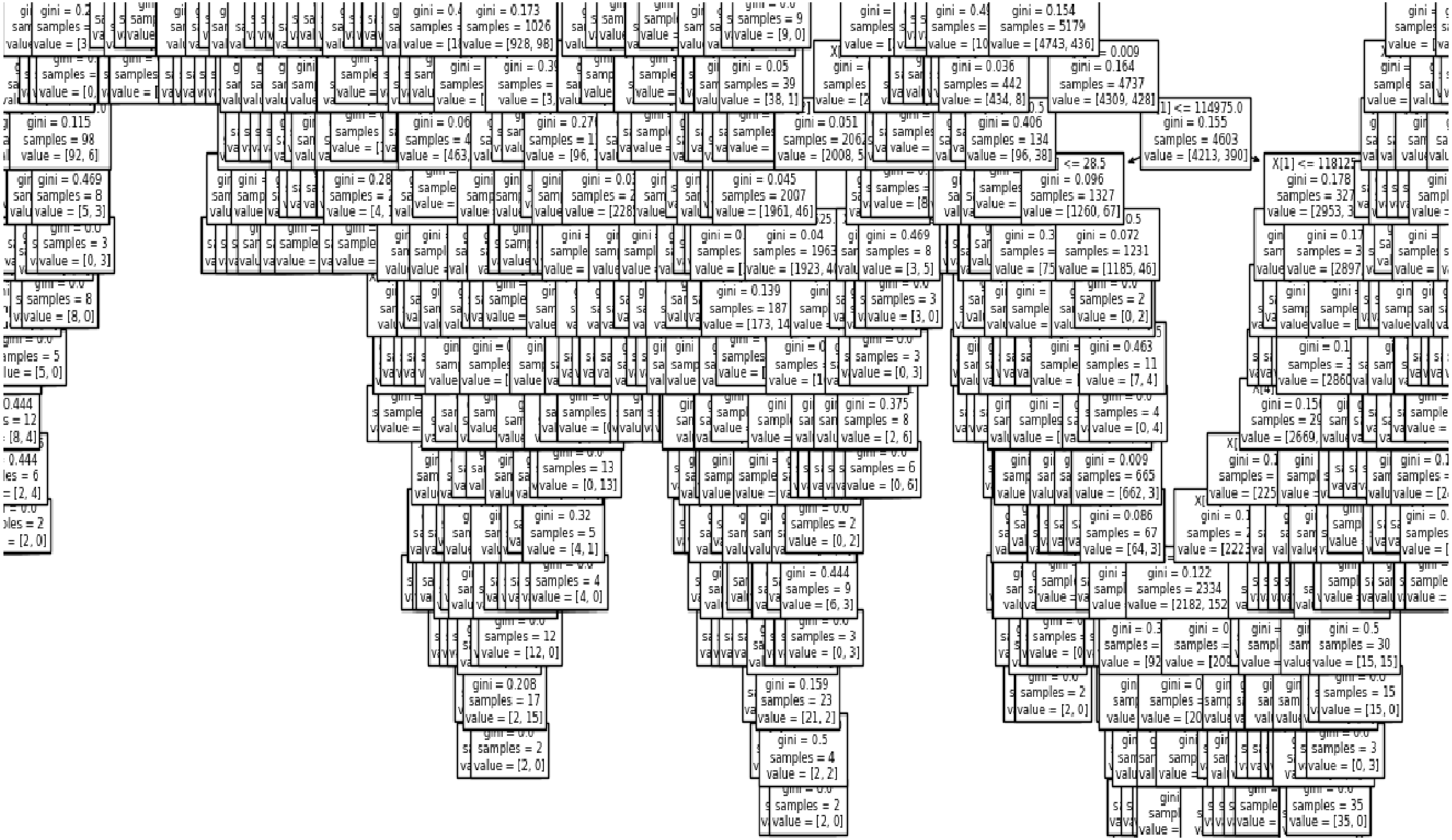


Figure D.2 – Second part of Decision tree

Continuation of appendix D

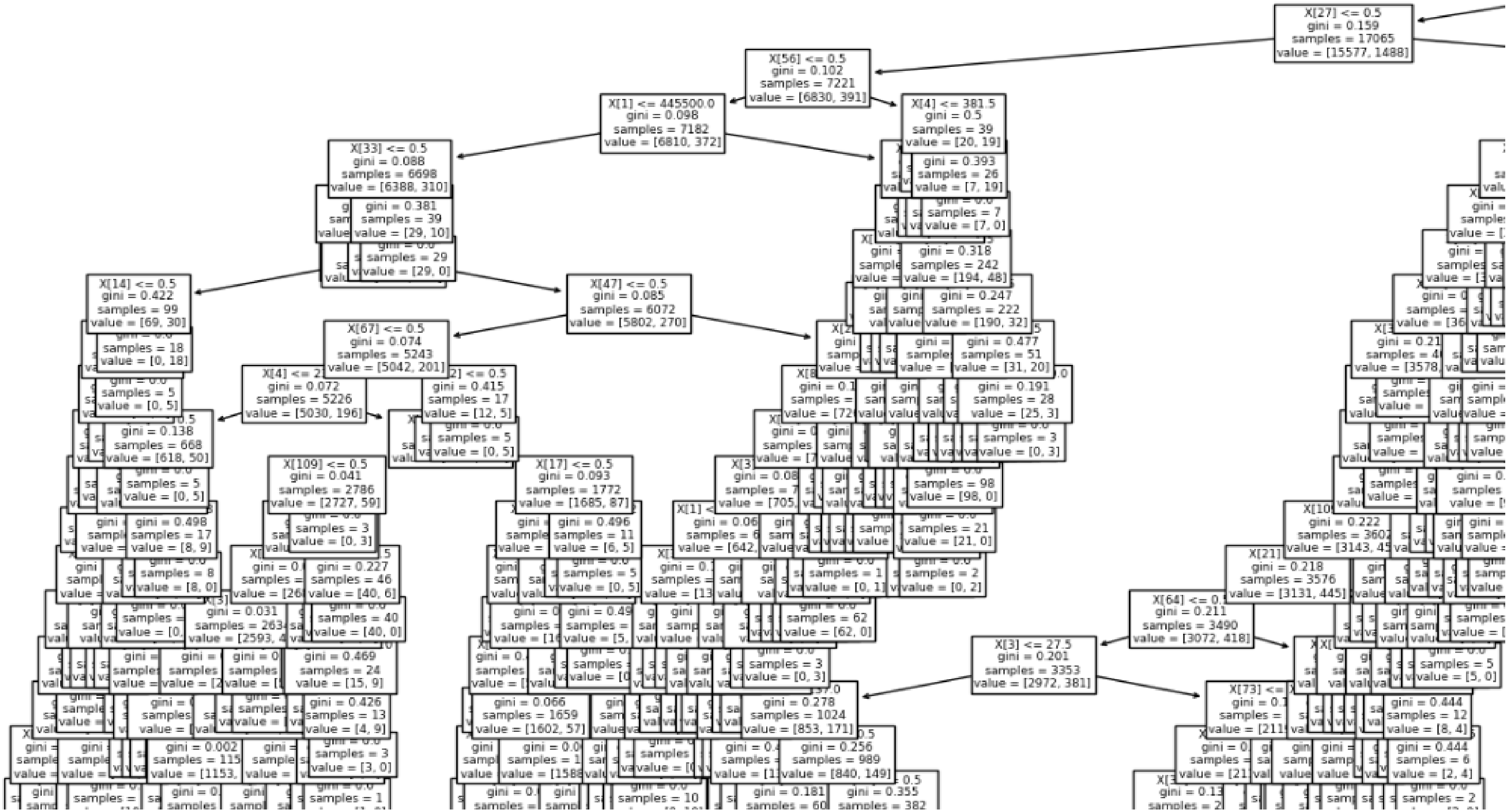


Figure D.3 – Third part of Decision tree

Continuation of appendix D

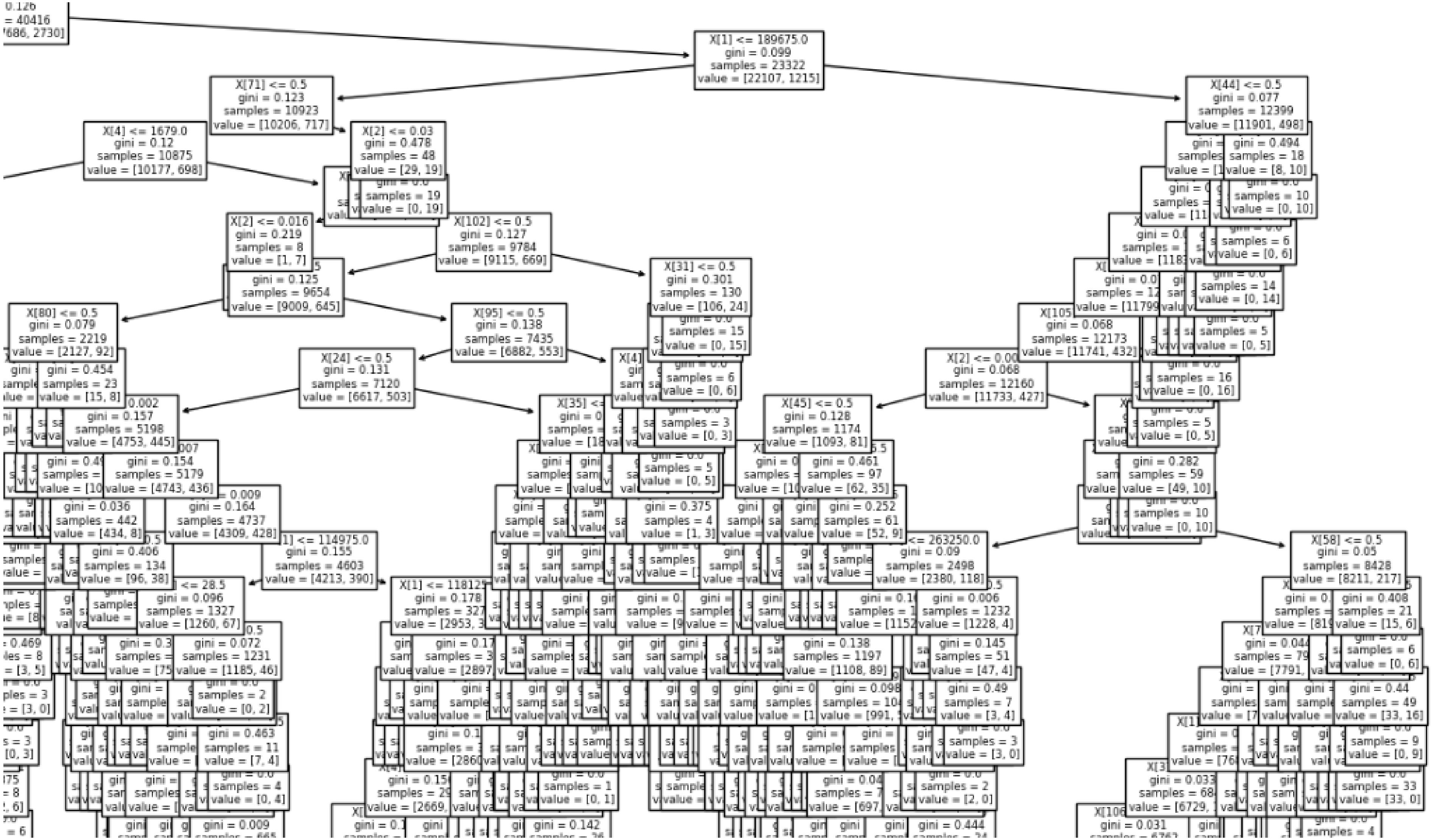


Figure D.4 – Fourth part of Decision tree