TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Research on Credit Risk of Bank Lending in the Context of Big Data (Исследование кредитного риска банковского кредитования в контексте больших данных)

УДК 004.65:004.451:336.774:005.032.2

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И	Цзинь Юйбо		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Е. В.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		

Томск – 2022

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований

ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Савельев А.О.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ПМ0И	Цзинь Юйбо

Тема работы:

Research on Credit Risk of Bank Lending in the Context of Big Data (Исследование кредитного риска банковского кредитования в контексте больших данных)	
Утверждена приказом директора (дата, номер)	№ 45-47/с от 14.02.2022

Срок сдачи студентом выполненной работы:	15.06.2022

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	В работе рассматриваются основные риски, с которыми сталкиваются банки. Одним из них является кредитный риск, основным элементом которого является заем. Эта работа помогает выявить хороших кредитных клиентов с помощью анализа больших данных в соответствии с личной информацией, чтобы выявить хорошего и плохого клиента.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1.Прочитайте и предварительно обработайте необработанный набор данных о клиентах банка. 2.Возобновите модель логистической регрессии и нарисуйте кривую ROC. 3.Используйте алгоритм случайного леса и алгоритм дерева решений для моделирования и прогнозирования. 4.Сравните и просмотрите важность каждой переменной. 5.Сравните точность разных моделей. 6.Рассчитайте информационное значение каждой переменной. 7.Выберите последнюю подходящую переменную и выполните обработку биннинга. 8.Разработайте систему показателей и визуализацию графического интерфейса пользователя. 9.Работа над разделом по финансовому менеджменту. 10.Работа над разделом по социальной ответственности.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Скриншот программы. 2. UML диаграммы. 3. Диаграмма графического интерфейса. 4. Диаграмма Ганта.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к.ф.-м.н., доцент Губин Е. И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.ф.н., доцент Меньшикова Е. В.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2022

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н., доцент		1.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ0И	Цзинь Юйбо		1.03.2022

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2021 /2022 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	15.06.2022

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2022	Основная часть	70
10.06.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2022	Социальная ответственность	10
10.06.2022	Английский язык	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

TASK FOR SECTION
«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING»

To the student:

Group	Full name
8PM0I	Yubo Jin

School	Information Tech & Robotics	**Division**	Big Data Solution
Degree	Master	**Educational Program**	09.04.04.Software Engineering

Title of graduation thesis:

Research on Credit Risk of Bank Lending in the Context of Big Data	
Input data to the section «Financial management, resource efficiency and resource saving»:	
1. *Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human*	– Salary costs – 272740.00 – STR budget – 157,504.54
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs – 5,8 rub per 1 kW
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Labor tax – 27,1 %; – Overhead costs – 30%;
The list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	– comparative analysis with other researches in this field;
2. *Development of charter for scientific-research project*	– SWOT-analysis;
3. *Scheduling of STR management process: structure and timeline, budget, risk management*	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. *Resource efficiency*	– integral indicator of resource efficiency for the developed project.
A list of graphic material *(with list of mandatory blueprints):*	
1. *Competitiveness analysis* 2. *SWOT- analysis* 3. *Gantt chart and budget of scientific research* 4. *Assessment of resource, financial and economic efficiency of STR* 5. *Potential risks*	

Date of issue of the task for the section according to the schedule	

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	E.V. Menshikova	PhD		

The task was accepted by the student:

Group	Full name	Signature	Date
8PM0I	Yubo Jin		

TASK FOR CHAPTER
«SOCIAL RESPONSIBILITY»

Student:

Group	Name		
8PM0I	Yubo Jin		
School	Information Tech & Robotics	**Division**	Big Data Solution
Educational level	Master degree	**Course/Specialty**	09.04.04. Software Engineering

Topic of FQW:

Research on Credit Risk of Bank Lending in the Context of Big Data	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	– Screen credit customers, according to the personal information on the finance dataset. – Model and evaluate accuracy using random forest and decision tree models.
List of questions to be researched, designed and developed:	
1. Legal and organizational issues of occupational safety – consider special (specific to the projected work area) law norms of labor legislation. – indicate the features of the labor legislation in relation to the specific conditions of the project.	– GOST 12.2.032-78 SSBT. Workplace when performing work while sitting General ergonomic requirements. – SanPiN 2.2.2/2.4.1340-03. Hygienic requirements for personal electronic computers and organization of work.
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the source of factor, the impact on human's body 2.2. Suggest measures to reduce the impact of identified harmful and dangerous factors	– Lack or lack of natural light, insufficient illumination. – Increased voltage in an electrical circuit, the closure of which can pass through the human body. – Physical overload (static-long-term preservation of a certain posture).
3. Environmental Safety: Influence on the atmosphere, hydrosphere, lithosphere	– Hydrosphere: Computer components that contain hazardous materials: lead, cadmium, lithium, alkaline manganese, and mercury. – Lithosphere: Computer components that contain plastic, glass, lead, barium, and rare earth metals.
4. Emergency Safety: Describe the most likely emergency situation	– Fire
Date issue of the task for the chapter	

Consultant:

Post	Name	Academic degree	Date	Signature
Docent professor	Antonevich O. A	PhD		

Student:

Group	Name	Date	Signature
8PM0I	Yubo Jin		

Abstract

The work contains an explanatory note on 102 sheets, contains 54 figures, 20 tables, 1 application.

Key words: finance dataset, correlation, binning, information value, logistic regression, random forest, visual scoring calculator.

In the field of financial risk control, everyone should know the scorecard. The so-called scorecard is to score the credit customers, the model gets the score, and the evaluation result is given by setting the threshold. The result can be directly used for passing or rejecting, or for policy application.

One of the primary risks faced by banks is credit risk, of which loan risk is the main element. This thesis aims to screen credit customers with big data analysis (logistic regression), according to the personal information on the finance dataset, find the right person (good or bad).

This work aims to screen credit customers according to the personal information on the finance dataset, finding the right person (the good customer). A bank credit score card is established for the bank's customer data through GUI, and a score can be obtained by entering some customer information, which provides help for the bank's customer classification.

CONTENT

1. Project Introduction

Data mining is the process of extracting information and knowledge that people do not know in advance but that has potential usefulness from a large amount of incomplete, noisy, fuzzy, and random actual data. Generally speaking, the results of data mining do not require completely accurate knowledge, but rather a general trend. As far as specific applications are concerned, data mining is the process of using various analytical tools to discover relationships between models and data in massive data sets, and these models and relationships can be used to make predictions. One of the primary risks faced by banks is credit risk, of which loan risk is the main element.

Typical bank big data application scenarios focus on database marketing, user management, data risk control, product design, and decision support. At present, the commercial application of big data in banks is mainly based on their own transaction data and customer data, external data is supplemented by descriptive data analysis, predictive data modeling is supplemented, and business customers are the main business. Supplemented by operating products.

Most of the banking data is structured data with strong financial attributes, which are stored in traditional relational databases and data warehouses. Through data mining, some of the knowledge hidden in transaction data with commercial value can be analyzed.

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

This work aims to screen credit customers according to the personal information on the finance dataset, finding the right person (the good customer), including:

- Data reading and preprocessing
- Correlation analysis of variables

- Model building
- Parameter optimization
- Model prediction and feature importance
- Visualize decision trees and random forests
- Compare the classification accuracy of decision trees and random forests.
- Make a visual scorecard

2. Dataset introduction:

The data is the personal information of three thousand users selected from a bank.The purpose is to build a model based on the existing customer information and default performance to predict whether to loan.(good or bad)

Table 1. Dataset basic info

Variable Name	Description	Type
AGE	Age	integer
BUREAU	Credit Bureau Risk Class	integer
CAR	Type of Vehicle	string
CARDS	Credit Cards	string
CASH	Requested cash	string
CHILDREN	Num of Children	integer
DIV	Large region	integer
EC_CARD	EC_card holders	integer
FINLOAN	Num finished Loans	integer
GB	Good/Bad	G/B
INC	Salary	integer
INC1	Salary+ec_card	integer
INCOME	Income	integer
LOANS	Num of running loans	integer
LOCATION	Location of Credit Bureau	integer
NAT	Nationality	string
NMBLOAN	Num Mybank Loans	integer
PERS_H	Num in Household	integer
PRODUCT	Type of Business	string
PROF	Profession	string

REGN	Region	integer
RESID	Residence Type	string
STATUS	Status	integer
TEL	Telephone	integer
TITLE	Title	integer
TMADD	Time at Address	integer
TMJOB1	Time at Job	integer

3. Data preprocessing

3.1 Data preprocessing by SAS

3.1.1 Import data

The first step is data import. By using PROC IMPORT I import our dataset and the file I used here is (.xls) file.

This is my code:

```
proc import
 datafile='C:\Users\Administrator\Desktop\sas finance\accepts.xls'
 out=yubo.test1
 replace
 dbms=xls;
 getnames=yes;
 run;
```

	TITLE	CHILDREN	PERS_H	AGE	TMADD	TMJOB1	TEL	NMBLOAN	FIN
1	R	2	4	39	9	30	2	2	1
2	R	3	5	28	0	168	1	0	0
3	H	0	2	22	3	84	2	0	1
4	R	0	2	31	36	144	2	0	1
5	R	0	2	28	36	33	2	0	0
6	R	0	2	23	9	9	2	0	0
7	R	0	2	29	9	54	2	2	1
8	H	0	2	25	12	27	2	2	1
9	R	0	1	46	144	30	1	0	0
10	H	2	3	38	168	27	1	0	0
11	H	1	2	43	42	144	1	0	0
12	R	2	3	29	30	144	1	0	0

Figure 1. Import data by SAS

3.1.2 View Descriptive statistics of variables

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

This part is about descriptive statistics.

Code:

```
proc means data=yubo.test1
n min max mean median std maxdec=2;
title 'Descriptive statistics of variables(n min max mean median std)';
run;
```

Descriptive statistics of variables(n min max mean median std)

The MEANS Procedure

Variable	Label	N	Minimum	Maximum	Mean	Median	Std Dev
CHILDREN	CHILDREN	3000	0.00	23.00	0.82	0.00	1.12
PERS_H	PERS_H	3000	1.00	25.00	2.36	2.00	1.42
AGE	AGE	3000	18.00	71.00	34.05	31.00	10.95
TMADD	TMADD	3000	0.00	999.00	119.28	60.00	180.09
TMJOB1	TMJOB1	3000	0.00	999.00	79.43	39.00	124.27
TEL	TEL	3000	0.00	2.00	1.82	2.00	0.39
NMBLOAN	NMBLOAN	3000	0.00	2.00	0.58	0.00	0.89
FINLOAN	FINLOAN	3000	0.00	1.00	0.48	0.00	0.50
INCOME	INCOME	3000	0.00	100000.00	1996.80	2100.00	2318.52
EC_CARD	EC_CARD	3000	0.00	1.00	0.26	0.00	0.44
INC	INC	3000	0.00	100000.00	31095.83	2500.00	45055.14
INC1	INC1	3000	0.00	5.00	2.39	2.00	1.21
BUREAU	BUREAU	3000	1.00	3.00	1.68	1.00	0.95
LOANS	LOANS	3000	0.00	9.00	1.01	1.00	1.11
REGN	REGN	3000	0.00	9.00	3.29	4.00	2.55
CASH	CASH	3000	0.00	100000.00	2497.13	1400.00	6360.28
GB	GB	3000	0.00	1.00	0.50	0.50	0.50

Figure 2. Descriptive statistics of variables by SAS

3.1.3 Fill in missing values

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data.

Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present.

- Fewer missing values: Directly delete samples with missing values.
- Moderate Missing Values: Impute missing values based on correlations between variables.
- More missing values: List feature values as attributes

Let us find missing values.

Code:

```
PROC FORMAT;
VALUE $MISSFMT ' ' ='MISSING' OTHER='NOT MISSING';
VALUE MISSFMT . ='MISSING' OTHER='NOT MISSING';
RUN;
PROC FREQ DATA=yubo.test1;
FORMAT _CHAR_ $MISSFMT.;
TABLES _CHAR_ /MISSING MISSPRINT NOCUM NOPERCENT;
FORMAT _NUMERIC_ MISSFMT.;
TABLES _NUMERIC_ /MISSING MISSPRINT NOCUM NOPERCENT;
RUN;
```

The FREQ Procedure

TITLE	
TITLE	Frequency
NOT MISSING	3000

PRODUCT	
PRODUCT	Frequency
MISSING	12
NOT MISSING	2988

RESID	
RESID	Frequency
MISSING	535
NOT MISSING	2465

NAT	
NAT	Frequency
NOT MISSING	3000

PROF	
PROF	Frequency
MISSING	1
NOT MISSING	2999

CAR	
CAR	Frequency
NOT MISSING	3000

CARDS	
CARDS	Frequency
NOT MISSING	3000

CHILDREN	
CHILDREN	Frequency
NOT MISSING	3000

PERS_H	
PERS_H	Frequency
NOT MISSING	3000

Figure 3. Frequency of dataset by SAS

Yes, you have seen it write these numbers is my missing values in each column. So, I change missing to most common value.

3.1.4 Handle outliers values

Code:

```
data Work.test2;
set Work.test1;
if product='' then product='Radio, TV, Hifi';
if resid='' then resid='Lease';
if prof='' then prof='Others';
run;
```

```
find outliers
proc univariate data=Work.test2 robustscale plot;
var inc age income;
run;
```

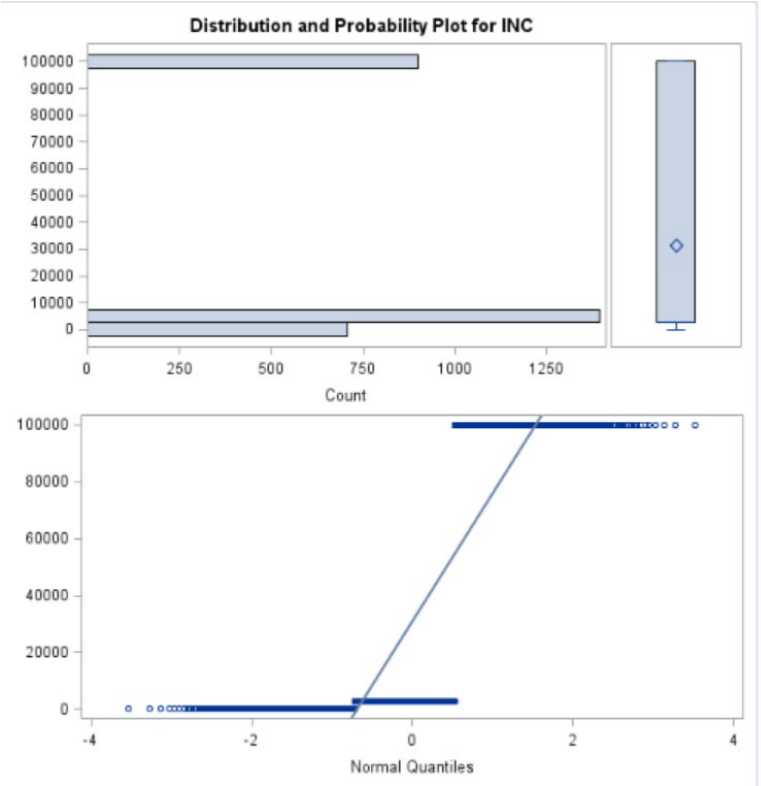


Figure 4. Outliers of inc variables by SAS

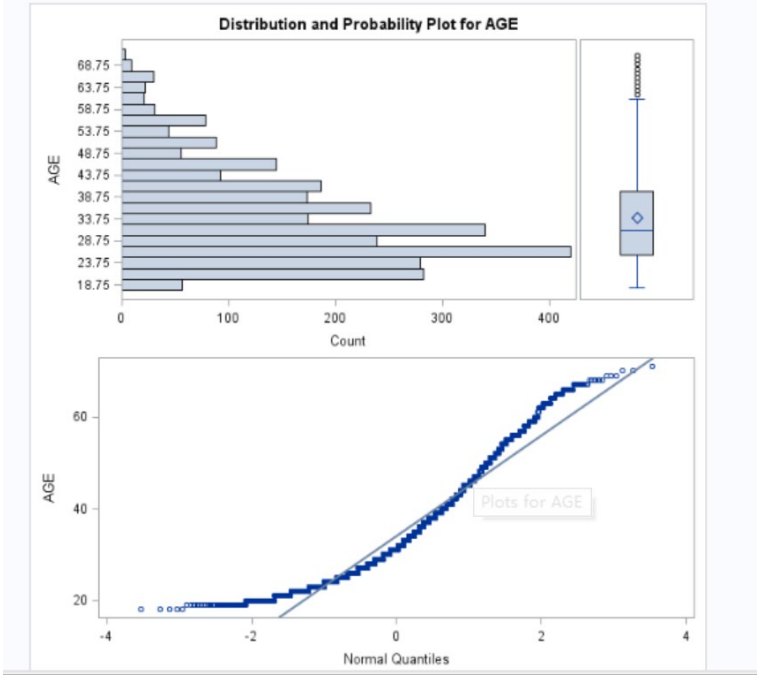


Figure 5. Outliers of age variables by SAS

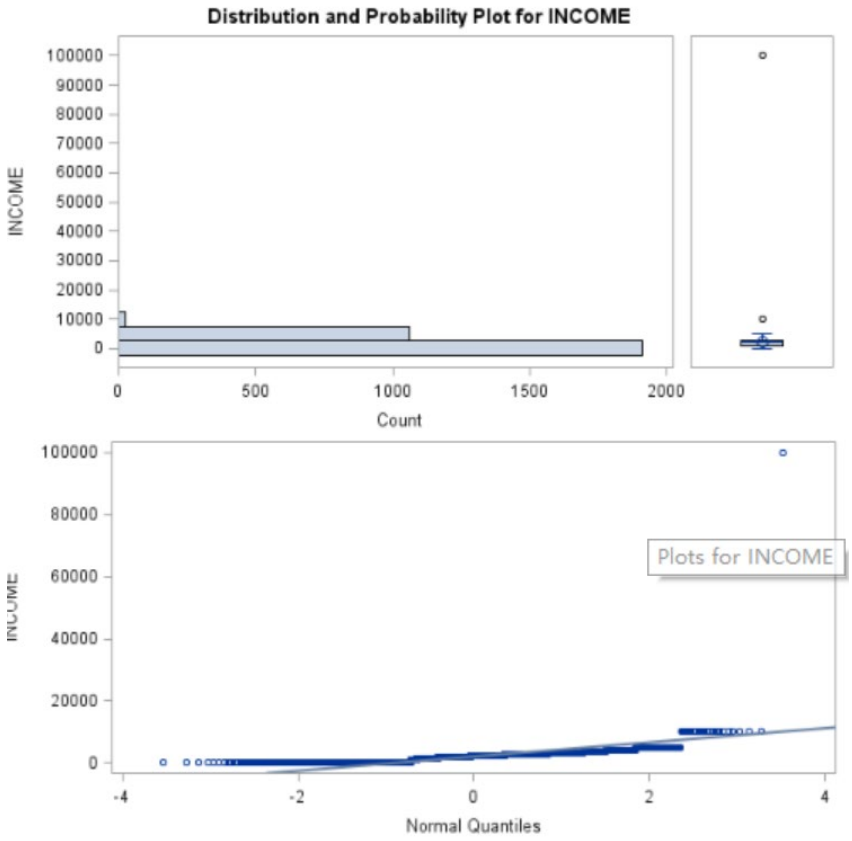


Figure 6. Outliers of income variables by SAS

So, from these pictures, we can get the outliers of age, inc and income, then I will delete outliers.

Code:

```
data Work.test2;
set Work.test1;
if children=23 then delete;
if income=100000 then delete;
if cash=100000 then delete;
run;
```

3.1.5 Drop duplicates

This part is about duplicates.

Code:

```
proc sort data=Work.test2
nodupkey out=NotDuplicate;
by title tel nmbloan finloan income ec_card inc1
children pers_h age tmadd tmjob1 bureau loans regn
cash product resid nat prof car cards;
run;
```

```
NOTE: There were 2987 observations read from the data set YUBO.TEST2.
NOTE: 0 observations with duplicate key values were deleted.
NOTE: The data set WORK.NOTDUPLICATE has 2987 observations and 24 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

Figure 7. Handling results for duplicate values by SAS

3.1.6 Digitalization of data

This is about digitalization of data.

Code:

```
data Work.test3;
set Work.test2;
if title='H'   then title_=1;
if title='R'   then title_=2;

if product='Cars'                then product_=1;
if product='Dept. Store,Mail'    then product_=2;
if product='Radio, TV, Hifi'     then product_=3;
if product='Leisure'             then product_=4;
if product='Others'              then product_=5;
```

```
if product='Furniture,Carpet'    then product_=6;

if resid='Lease'                 then resid_=1;
if resid='Owner'                 then resid_=2;

if nat='German'                  then nat_=1;
if nat='Greek'                   then nat_=2;
if nat='Italian'                 then nat_=3;
if nat='Other European'          then nat_=4;
if nat='Turkish'                 then nat_=5;
if nat='Spanish/Portugue'        then nat_=6;
if nat='Others'                  then nat_=7;
if nat='Yugoslav'                then nat_=8;


if prof='State,Steel Ind,'       then prof_=1;
if prof='Civil Service, M'       then prof_=2;
if prof='Food,Building,Ca'       then prof_=3;
if prof='Military Service'       then prof_=4;
if prof='Others'                 then prof_=5;
if prof='Pensioner'              then prof_=6;
if prof='Sea Vojage, Gast'       then prof_=7;
if prof='Self-employed pe'       then prof_=8;
if prof='Chemical Industr'       then prof_=9;

if car='Car'                    then car_=1;
if car='Car and Motor bi'       then car_=2;
if car='Without Vehicle'        then car_=3;

if cards='American Express'      then cards_=1;
```

```
if cards='Cheque card'          then cards_=2;
if cards='Mastercard/Euroc'     then cards_=3;
if cards='Other credit car'     then cards_=4;
if cards='VISA Others'          then cards_=5;
if cards='VISA mybank'          then cards_=6;
if cards='no credit cards'      then cards_=7;

drop title product resid nat prof car cards;
run;
```

title_	product_	resid_	nat_	prof_	car_	cards_
2	6	1	1	2	1	7
2	6	1	1	2	3	7
1	1	1	1	5	1	7
2	6	1	1	5	3	7
2	6	2	1	5	1	7
2	1	1	5	5	1	7
2	6	1	1	5	3	7
1	3	1	1	5	1	7
2	3	1	1	5	3	7
1	3	1	1	5	1	7
1	4	1	1	5	3	7
2	6	1	1	2	1	7
2	6	1	1	5	1	7
1	6	1	1	6	1	2
2	3	.	1	5	3	7

Figure 8. Digitalization of data by SAS

3.2 Data preprocessing by Python

3.2.1 Import dataset and check it

The data is the personal information of three thousand users selected from a b ank

Columns: (['TITLE', 'CHILDREN', 'PERS_H', 'AGE', 'TMADD', 'TMJOB1', ' TEL',

'NMBLOAN', 'FINLOAN', 'INCOME', 'EC_CARD', 'INC', 'INC1', 'BURE AU',

'LOANS', 'REGN', 'CASH', 'PRODUCT', 'RESID', 'NAT', 'PROF', 'CAR',

'CARDS', 'GB'], dtype='object')

Code:

```
df = pd.read_csv(r'C:\Users\Administrator\Desktop\scientific_work\1.csv')
df.shape
df.columns
df.head(5)
```

```
[2]: # import data
     df = pd.read_csv(r'C:\Users\Administrator\Desktop\scientific_work\1.csv')

     ...

[4]: df.shape

[4]: (3000, 24)

[5]: df.columns

[5]: Index(['TITLE', 'CHILDREN', 'PERS_H', 'AGE', 'TMADD', 'TMJOB1', 'TEL',
            'NMBLOAN', 'FINLOAN', 'INCOME', 'EC_CARD', 'INC', 'INC1', 'BUREAU',
            'LOANS', 'REGN', 'CASH', 'PRODUCT', 'RESID', 'NAT', 'PROF', 'CAR',
            'CARDS', 'GB'],
           dtype='object')
```

Figure 9. Import and check the shape of dataset by Python

```
]: df.head(5)
```

]:

	TITLE	CHILDREN	PERS_H	AGE	TMADD	TMJOB1	TEL	NMBLOAN	FINLOAN	INCOME	...	LOANS	REGN	CASH	PRODUCT	RESID	NAT	PROF	CAR	CARDS	GB
0	R	2	4	39	9	30	2	2	1	2200	...	2	0	4000	Furniture,Carpet	Lease	German	Civil Service, M	Car	no credit cards	1
1	R	3	5	28	0	168	1	0	0	1700	...	1	0	4000	Furniture,Carpet	Lease	German	Civil Service, M	Without Vehicle	no credit cards	1
2	H	0	2	22	3	84	2	0	1	2500	...	1	8	1100	Cars	Lease	German	Others	Car	no credit cards	1
3	R	0	2	31	36	144	2	0	1	2000	...	1	2	2000	Furniture,Carpet	Lease	German	Others	Without Vehicle	no credit cards	1
4	R	0	2	28	36	33	2	0	0	3500	...	4	7	15000	Furniture,Carpet	Owner	German	Others	Car	no credit cards	1

5 rows × 24 columns

Figure 10. Check part of dataset by Python

3.2.2 View Descriptive statistics of variables

Code:

```
df.describe().T
```

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
CHILDREN	3000.0	0.819667	1.122007	0.0	0.00	0.0	1.0	23.0
PERS_H	3000.0	2.358667	1.416814	1.0	1.00	2.0	3.0	25.0
AGE	3000.0	34.047667	10.951501	18.0	25.75	31.0	40.0	71.0
TMADD	3000.0	119.282000	180.089142	0.0	21.00	60.0	168.0	999.0
TMJOB1	3000.0	79.431000	124.272858	0.0	18.00	39.0	96.0	999.0
TEL	3000.0	1.815000	0.389220	0.0	2.00	2.0	2.0	2.0
NMBLOAN	3000.0	0.580667	0.887932	0.0	0.00	0.0	2.0	2.0
FINLOAN	3000.0	0.477000	0.499554	0.0	0.00	0.0	1.0	1.0
INCOME	3000.0	1996.800000	2318.521427	0.0	1000.00	2100.0	2700.0	100000.0
EC_CARD	3000.0	0.258333	0.437791	0.0	0.00	0.0	1.0	1.0
INC	3000.0	31095.833333	45055.137196	0.0	2500.00	2500.0	100000.0	100000.0
INC1	3000.0	2.385667	1.211370	0.0	2.00	2.0	4.0	5.0
BUREAU	3000.0	1.683333	0.946584	1.0	1.00	1.0	3.0	3.0
LOANS	3000.0	1.008000	1.112512	0.0	0.00	1.0	2.0	9.0
REGN	3000.0	3.293667	2.551914	0.0	0.00	4.0	5.0	9.0
CASH	3000.0	2497.133333	6360.283237	0.0	900.00	1400.0	2500.0	100000.0
GB	3000.0	0.500000	0.500083	0.0	0.00	0.5	1.0	1.0

Figure 11. Descriptive statistics of variables by Python

3.2.3 Fill in missing values and drop duplicates

Methods for dealing with missing values include the following:

Fewer missing values: Directly delete samples with missing values.

Moderate Missing Values: Impute missing values based on correlations between variables.

More missing values: List feature values as attributes.

So first check our dataset for missing cases.

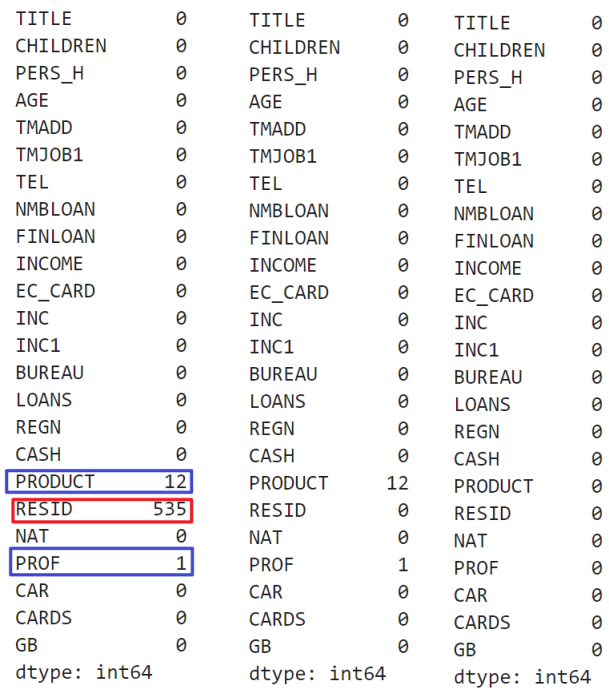


Figure 12. Dealing with missing values by Python

From the above figure, we find that the missing rate of the variable RESID is relatively large, so we fill in the missing values according to the correlation between the variables, but the data type of RESID is a string, so we use the mode to fill in here. And the variable PRODUCT and PROF has fewer missing values, so we delete it directly.

After dealing with missing values, we review and drop duplicates.

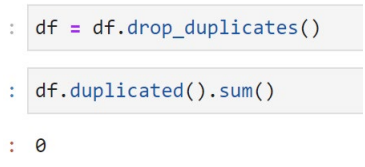


Figure 13. Review and drop duplicates by Python

Code:

```
RESID=df.loc[:,"RESID"].values.reshape(-1,1)
imp_mode=SimpleImputer(strategy="most_frequent")
df.loc[:,"RESID"]=imp_mode.fit_transform(RESID)
df = df.drop_duplicates()
df.isnull().sum()
df = df.dropna()
df.isnull().sum()
```

3.2.4 Handle outliers values

After dealing with missing values, deal with outliers. Outliers generally refer to values that deviate significantly from the data. In statistics, for example, outliers are defined as values less than Q1-1.5IQR or greater than Q3+1.5IQR. We observe and deal with outliers of each variable by drawing boxplots.

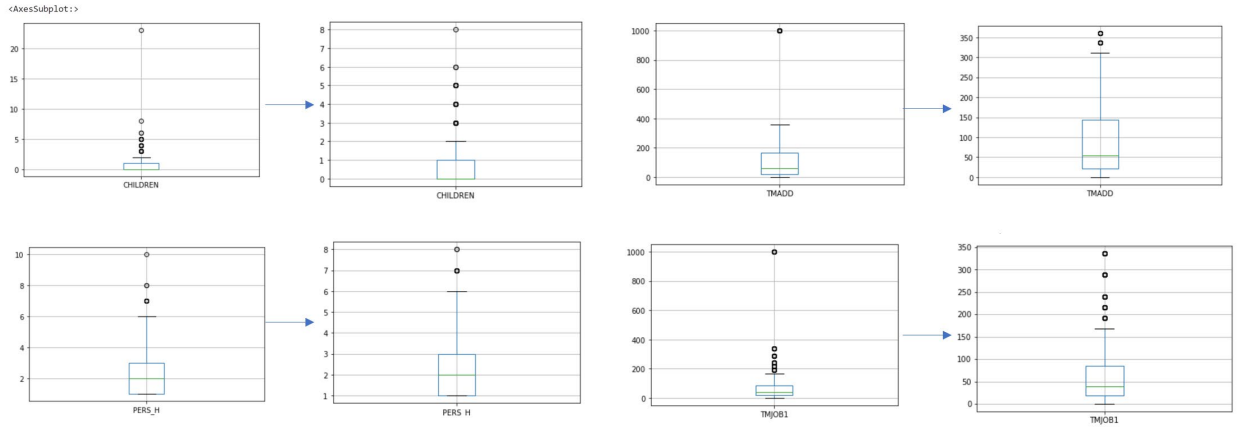


Figure 14. Handling outliers1 by Python

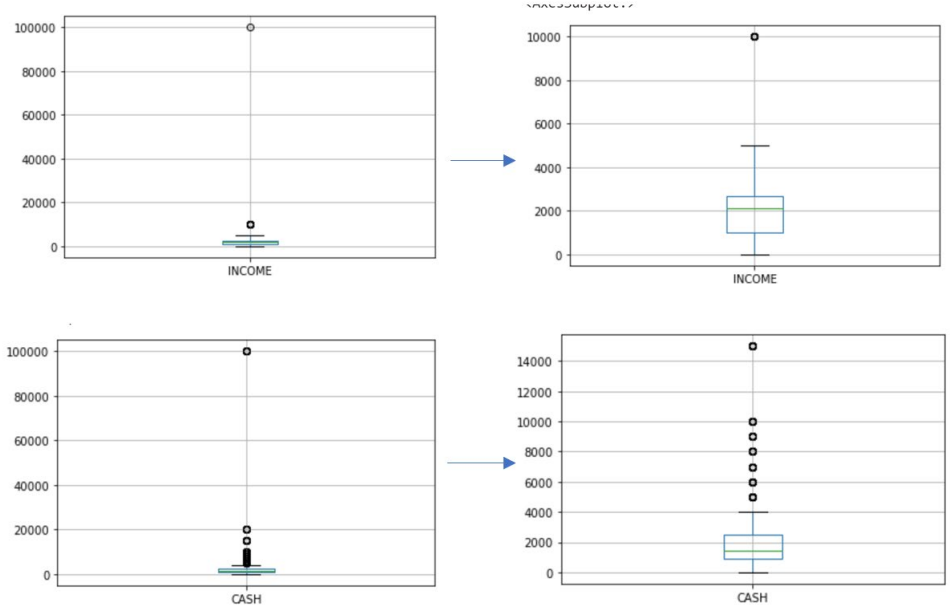


Figure 15. Handling outliers2 by Python

Code:

```
df[['CHILDREN']].boxplot()
df = df[df['CHILDREN']<10]
```

```
df[['CHILDREN']].boxplot()
df[['PERS_H']].boxplot()
df[['AGE']].boxplot()
df[['TMADD']].boxplot()
df = df[df['TMADD']<400]
df[['TMADD']].boxplot()
df[['TMJOB1']].boxplot()
df = df[df['TMJOB1']<400]
df[['TMJOB1']].boxplot()
df[['TEL']].boxplot()
df[['NMBLOAN']].boxplot()
df[['FINLOAN']].boxplot()
df[['INCOME']].boxplot()
df = df[df['INCOME']<20000]
df[['INCOME']].boxplot()
df[['EC_CARD']].boxplot()
df[['INC']].boxplot()
df[['INC1']].boxplot()
df[['BUREAU']].boxplot()
df[['LOANS']].boxplot()
df[['REGN']].boxplot()
df[['CASH']].boxplot()
df = df[df['CASH']<20000]
df[['CASH']].boxplot()
df[['GB']].boxplot()
```

3.2.5 Categorical Variable Handling

In general, common processing methods for categorical variables with small number of categories are: one-hot encoding, dummy encoding and label encoding. Before, when using SAS software to preprocess the data, we used the label encoding method. Label encoding is the serialized label encoding, which marks 0, 1, 2, ..., n according to the order and number of categorical variables. For easy comparison, we will use one-hot encoding for processing in the next python.

One-hot encoding, similar to dummy variables, is a way to convert categorical variables into several binary columns. where 1 means that an input belongs to that category.

```
df_dummies = pd.get_dummies(df)
```

```
df_dummies.head(10)
```

	CHILDREN	PERS_H	AGE	TMADD	TMJOB1	TEL	NMBLOAN	FINLOAN	INCOME	EC_CARD	...	CAR_Car	CAR_Car and Motor bi	CAR_Without Vehicle
0	2	4	39	9	30	2	2	1	2200	0	...	1	0	0
1	3	5	28	0	168	1	0	0	1700	0	...	0	0	1
2	0	2	22	3	84	2	0	1	2500	0	...	1	0	0
3	0	2	31	36	144	2	0	1	2000	0	...	0	0	1
4	0	2	28	36	33	2	0	0	3500	0	...	1	0	0
5	0	2	23	9	9	2	0	0	1900	0	...	1	0	0
6	0	2	29	9	54	2	2	1	1800	0	...	0	0	1
7	0	2	25	12	27	2	2	1	5000	0	...	1	0	0
8	0	1	46	144	30	1	0	0	1500	0	...	0	0	1
9	2	3	38	168	27	1	0	0	3400	0	...	1	0	0

10 rows × 54 columns

Figure 16. One-hot encoding of dataset by Python

From a machine learning perspective, one-hot encoding is not a good way to encode categorical variables.

Because one-hot encoding adds a large number of dimensions, it is necessary to enumerate all the value cases under this feature.

One-hot encoding not only adds a lot of dimension to the dataset, but there is actually not much information, many times 1 is scattered among many zeros, that is, useful information is scattered in a large amount of data. This can lead to unusually sparse results, making it difficult to optimize, especially for neural networks.

Worse, there is a linear relationship between each informative sparse column. This means that one variable can easily be predicted using other variables, leading to problems of parallelism and multicollinearity in high dimensions.

The loss function of some models is sensitive to the size of the value, that is, the size of the value between the variables is relatively meaningful, such as logistic regression, SVM, etc., we temporarily call it a type A model; some models themselves are not sensitive to numerical changes. , the meaning of numerical existence is more for sorting, that is, there is no difference between 0.1, 0.2, 0.3 and 1, 2, and 3. Most of these models are tree models, which are temporarily called B-type models.

Type A model

Categorical variables must do one hot encoding, because label encoding has no numerical meaning.

But for variables with many categories, doing one hot encoding will make the generated variables too sparse, so here are some empirical methods. Another method is to place the top n categories with the most occurrences of one hot, and other categories in the variables of other categories; you can also use the positive rate in the y value (training value) to merge, but it is prone to overfitting.

Type B model

If the B-type model is used and it is an ordered variable, label encoding is preferred, and the assignment must be consistent with the order.

If it is an unordered variable, the two methods are not very different in many cases, but the effect of label encoding is generally better than one hot encoding in actual use. This is because in the tree model, label encoding can at least achieve the same effect as one hot encoding, and the extra information is that the value after label encoding itself has a sorting effect, which can play the effect of merging categorical variables. , this effect is more pronounced for variables with more categories.

3.2.6 EDA(Exploratory Data Analysis)

Before building a model, we generally perform Exploratory Data Analysis on the existing data. EDA refers to the exploration of existing data (especially raw data from surveys or observations) with as few a priori assumptions as possible. Commonly used exploratory data analysis methods are: histogram, scatter plot and boxplot. Here we will use histogram.

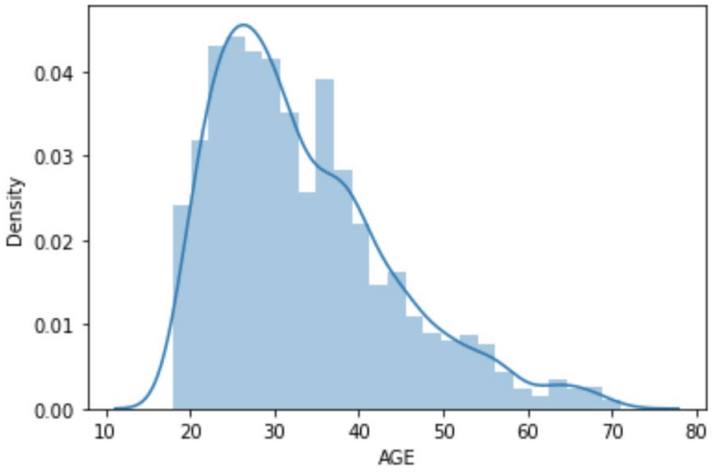


Figure 17. Histogram of AGE

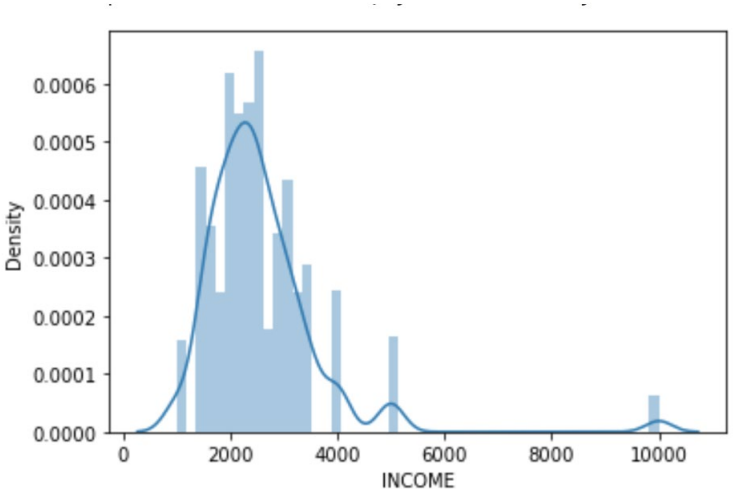


Figure 18. Histogram of INCOME

4. Feature variable selection

4.1 Correlation Matrix

4.1.1 Correlation Matrix (SAS)

Feature variable selection (ranking) is very important for data analysis, machine learning practitioners. Good feature selection can improve the performance of the model and help us understand the characteristics and underlying structure of the data, which plays an important role in further improving the model and algorithm.

Next, we got to the point where we started to calculate the correlation between the sets of data. We generally use the correlation coefficient to describe the correlation between the two sets of data, and the correlation coefficient is obtained by dividing the covariance by the standard deviation of the two variables, and the correlation coefficient will be between [-1, 1], - 1 means completely negative correlation and 1 means completely correlated.

Correlation Matrix

		CHILDREN	PERS_H	AGE	TMADD	TMJOB1	TEL	NMBLOAN	FINLOAN	INCOME	EC_CARD	INC	INC1	BUREAU	LOANS	REGN	CASH	GB
CHILDREN	CHILDREN	1.0000	0.9461	0.1150	-.0448	0.0700	0.0166	0.0461	0.0782	0.1486	0.0188	0.2599	0.1982	-.0857	0.1374	-.0398	0.0418	-.0768
PERS_H	PERS_H	0.9461	1.0000	0.1972	-.0399	0.1057	0.0379	0.0618	0.0954	0.1674	0.0265	0.2986	0.2253	-.1013	0.1640	-.0382	0.0622	-.1367
AGE	AGE	0.1150	0.1972	1.0000	0.2642	0.3594	-.0011	0.0153	0.0363	0.0187	0.0876	0.0829	0.0320	-.0288	0.0559	-.1377	0.0866	-.2747
TMADD	TMADD	-.0448	-.0399	0.2642	1.0000	0.1938	-.0542	-.0281	-.0346	-.0539	0.0574	-.0311	-.0331	0.0307	-.0215	-.1117	0.0167	-.0541
TMJOB1	TMJOB1	0.0700	0.1057	0.3594	0.1938	1.0000	0.0235	0.0744	0.0607	0.0695	0.0239	0.1400	0.0966	-.0641	0.0790	-.0385	0.0663	-.1530
TEL	TEL	0.0166	0.0379	-.0011	-.0542	0.0235	1.0000	0.0845	0.1539	0.1505	-.0879	0.1646	0.1537	-.0304	0.0666	0.2892	0.0226	-.1258
NMBLOAN	NMBLOAN	0.0461	0.0618	0.0153	-.0281	0.0744	0.0845	1.0000	0.4149	0.1189	-.0610	0.1302	0.1228	-.2457	0.2532	0.0704	-.0179	-.0767
FINLOAN	FINLOAN	0.0782	0.0954	0.0363	-.0346	0.0607	0.1539	0.4149	1.0000	0.1772	-.0592	0.2334	0.2064	-.2482	0.2744	0.1806	-.0221	-.0634
INCOME	INCOME	0.1486	0.1674	0.0187	-.0539	0.0695	0.1505	0.1189	0.1772	1.0000	-.5822	0.6846	0.8193	-.1240	0.1524	0.2752	0.1713	0.0681
EC_CARD	EC_CARD	0.0188	0.0265	0.0876	0.0574	0.0239	-.0879	-.0610	-.0592	-.5822	1.0000	-.2987	-.4380	0.0155	-.0009	-.2668	-.0941	-.1869
INC	INC	0.2599	0.2986	0.0829	-.0311	0.1400	0.1646	0.1302	0.2334	0.6846	-.2987	1.0000	0.9158	-.1445	0.2128	0.2077	0.1524	-.0493
INC1	INC1	0.1982	0.2253	0.0320	-.0331	0.0966	0.1537	0.1228	0.2064	0.8193	-.4380	0.9158	1.0000	-.1128	0.1728	0.2516	0.1522	0.0159
BUREAU	BUREAU	-.0857	-.1013	-.0288	0.0307	-.0641	-.0304	-.2457	-.2482	-.1240	0.0155	-.1445	-.1128	1.0000	-.6477	-.0311	-.0460	-.0245
LOANS	LOANS	0.1374	0.1640	0.0559	-.0215	0.0790	0.0666	0.2532	0.2744	0.1524	-.0009	0.2128	0.1728	-.6477	1.0000	0.0401	0.0630	0.0379
REGN	REGN	-.0398	-.0382	-.1377	-.1117	-.0385	0.2892	0.0704	0.1806	0.2752	-.2668	0.2077	0.2516	-.0311	0.0401	1.0000	0.0090	0.0406
CASH	CASH	0.0418	0.0622	0.0866	0.0167	0.0663	0.0226	-.0179	-.0221	0.1713	-.0941	0.1524	0.1522	-.0460	0.0630	0.0090	1.0000	-.0575
GB	GB	-.0768	-.1367	-.2747	-.0541	-.1530	-.1258	-.0767	-.0634	0.0681	-.1869	-.0493	0.0159	-.0245	0.0379	0.0406	-.0575	1.0000

Figure 19. Correlation Matrix

4.1.2 Correlation (Hive)

Hive is a data warehouse tool based on Hadoop, which can map structured data files to a database table and provide SQL-like query functions.

Code:

```
SELECT
        round(corr(INCOME,GB),2)                    as
correlation_INCOME_GB,
        round(corr(cash,GB),2) as correlation_CASH_GB,
        round(stddev_pop(AGE),2) as std_p,
        round(stddev_samp(AGE),2) as std_s,
        percentile(INCOME,0.75) as 75quartiles,
        percentile(INCOME,0.5) as median,
        percentile(INCOME,0.25) as 25quartiles,
        var_samp(CASH) as variance_s,
        var_pop(CASH) as variance_p
from finance;
```

	correlation_income_gb	correlation_cash_gb	std_p	std_s	75quartiles	median	25quartiles	variance_
1	0.06	-0.02	10.95	10.95	2700	2100	1000	40453202

variance_s	variance_p
40453202.84983891	40439718.448888965

Figure 20. Correlation and variance

4.1.3 Correlation (Python)

Let's first check the correlation between variables and have an intuitive understanding of the variables. We use the seaborn package in Python to call the heatmap() drawing function to draw the correlation diagram.

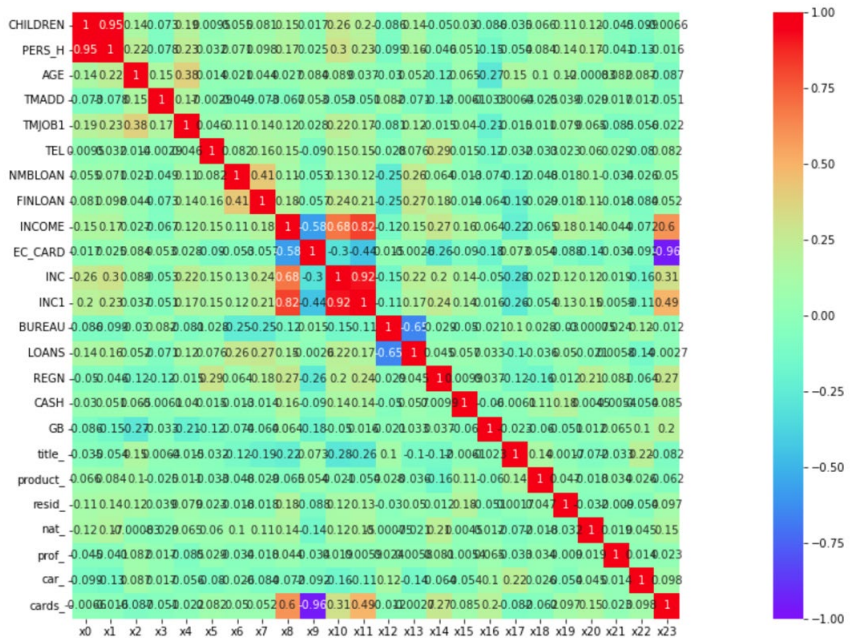


Figure 21. Correlation before processing

The correlation between most of the variables is very small, and there is no multicollinearity problem. A few of them have multicollinearity, that is, there may be two variables that are highly correlated and need to be eliminated.

('INCOME', 'EC_CARD', 'INC', 'INC1')=>('INCOME') and ('BUREAU', 'LOANS') => ('LOANS')

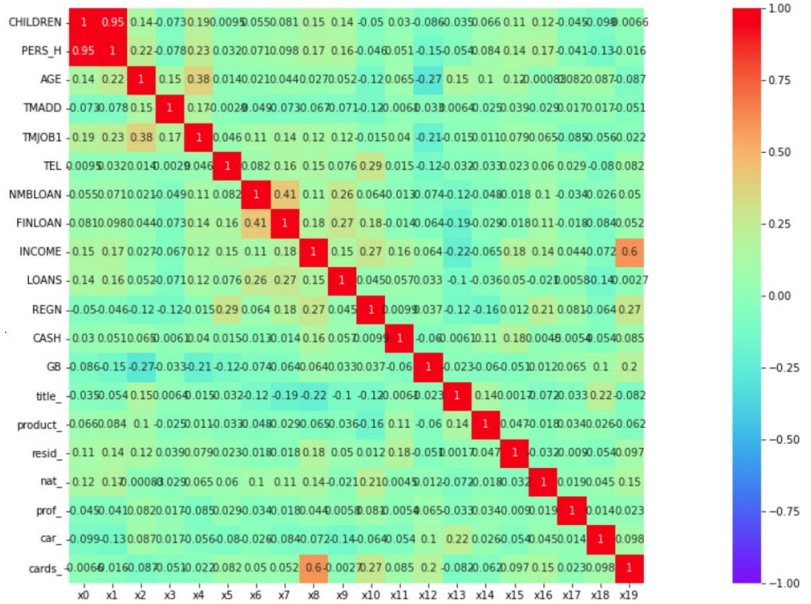


Figure 22. Processed correlation

4.2 Binning

Binning processing is also known as continuous variable discretization, there are generally equidistant, equal frequency, chi-square binning and other methods in the development of credit scorecards. First, the optimal segmentation of continuous variables is selected. When the distribution of continuous variables does not meet the requirements of optimal segmentation, the continuous variables are considered to be equally spaced. The code for optimal binning is as follows:

For AGE, INCOME and CASH TMJOB1 TMADD we use optimal binning to classify.

```
x2_d,x2_iv,x2_cut,x2_woe = mono_bin(y_trian,X_trian.AGE)
x3_d,x3_iv,x3_cut,x3_woe = mono_bin(y_trian,X_trian.TMADD)
x4_d,x4_iv,x4_cut,x4_woe = mono_bin(y_trian,X_trian.TMJOB1)
x8_d,x8_iv,x8_cut,x8_woe = mono_bin(y_trian,X_trian.INCOME)
x11_d,x11_iv,x11_cut,x11_woe = mono_bin(y_trian,X_trian.CASH)
```

Code:

```
def mono_bin(Y, X, n=10):
    r = 0
    good=Y.sum()
    bad=Y.count()-good
    while np.abs(r) < 1:
        d1 = pd.DataFrame({"X": X, "Y": Y, "Bucket":
pd.qcut(X, n,duplicates="drop")})
        d2 = d1.groupby('Bucket', as_index = True)
        r, p = stats.spearmanr(d2.mean().X,
d2.mean().Y)
        n = n - 1
    d3 = pd.DataFrame(d2.X.min(), columns = ['min'])
    d3['min']=d2.min().X
    d3['max'] = d2.max().X
    d3['sum'] = d2.sum().Y
    d3['total'] = d2.count().Y
```

```
        d3['rate'] = d2.mean().Y
        d3['woe']=np.log((d3['rate']/good)/((1-
d3['rate'])/bad))
        d3['goodattribute']=d3['sum']/good
        d3['badattribute']=(d3['total']-d3['sum'])/bad
        iv=((d3['goodattribute']-
d3['badattribute'])*d3['woe']).sum()
        d4 = (d3.sort_values(by =
'min')).reset_index(drop=True)
        woe=list(d4['woe'].round(3))
        cut=[]
        cut.append(float('-inf'))
        for i in range(1,n+1):
            qua=X.quantile(i/(n+1))
            cut.append(round(qua,4))
        cut.append(float('inf'))
        return d4,iv,cut,woe
    def self_bin(Y,X,cat):
        good=Y.sum()
        bad=Y.count()-good

d1=pd.DataFrame({'X':X,'Y':Y,'Bucket':pd.cut(X,cat)})
        d2=d1.groupby(['Bucket'])
        d3=pd.DataFrame(d2['X'].min(),columns=['min'])
        d3['min']=d2['X'].min()
        d3['max']=d2['X'].max()
        d3['sum']=d2['Y'].sum()
        d3['total']=d2['Y'].count()
        d3['rate']=d2['Y'].mean()
        d3['goodattribute']=d3['sum']/good
```

```
        d3['badattribute']=(d3['total']-d3['sum'])/bad

d3['woe']=np.log(d3['goodattribute']/d3['badattribute']
)
        iv=((d3['goodattribute']-
d3['badattribute'])*d3['woe']).sum()
        d4=d3.sort_values(by='min')
        print(d4)
        print('-'*40)
        woe=list(d3['woe'].values)
        return d4,iv,woe
```

However, other variables cannot be binned in this way, so we use manual selection: the selection method is as follows.

Code:

```
x0_cut=[ninf,0,1,2,5,pinf]
x1_cut=[ninf,1,2,3,5,pinf]
x5_cut=[ninf,1,2,pinf]
x6_cut = [ninf, 0, 1, 2, pinf]
x7_cut = [ninf, 0, 1, pinf]
x9_cut = [ninf, 0, 1, 2, pinf]
x10_cut = [ninf, 0, 4, 5, pinf]
x13_cut=[ninf,1,2,pinf]
x14_cut=[ninf,1,3,5,pinf]
x15_cut=[ninf,1,2,pinf]
x16_cut=[ninf,1,8,pinf]
x17_cut=[ninf,1,5,pinf]
x18_cut=[ninf,1,3,pinf]
x19_cut=[ninf,1,2,5,pinf]
```

```
             min  max  sum  total      rate  goodattribute  badattribute  \
Bucket
(-inf, 0.0]    0    0  843   1502  0.561252       0.583795      0.465724
(0.0, 1.0]     1    1  283    652  0.434049       0.195983      0.260777
(1.0, 2.0]     2    2  223    509  0.438114       0.154432      0.202120
(2.0, 5.0]     3    5   95    194  0.489691       0.065789      0.069965
(5.0, inf]     6    6    0      2  0.000000       0.000000      0.001413

                 woe
Bucket
(-inf, 0.0]  0.225956
(0.0, 1.0]  -0.285637
(1.0, 2.0]  -0.269108
(2.0, 5.0]  -0.061530
(5.0, inf]       -inf
---------------------------------------
             min  max  sum  total      rate  goodattribute  badattribute  \
Bucket
(-inf, 1.0]    1    1  679   1070  0.634579       0.470222      0.276325
(1.0, 2.0]     2    2  251    600  0.418333       0.173823      0.246643
(2.0, 3.0]     3    3  238    550  0.432727       0.164820      0.220495
(3.0, 5.0]     4    5  250    590  0.423729       0.173130      0.240283
(5.0, inf]     6    8   26     49  0.530612       0.018006      0.016254

                 woe
Bucket
(-inf, 1.0]  0.531626
(1.0, 2.0]  -0.349906
(2.0, 3.0]  -0.291020
(3.0, 5.0]  -0.327772
(5.0, inf]   0.102315
---------------------------------------
             min  max   sum  total      rate  goodattribute  badattribute  \
Bucket
(-inf, 1.0]  1.0  1.0   329    517  0.636364       0.227839      0.132862
(1.0, 2.0]   2.0  2.0  1115   2342  0.476089       0.772161      0.867138
(2.0, inf]   NaN  NaN     0      0       NaN       0.000000      0.000000
```

Figure 23. Partial binning results display

4.3 Information Value

When doing feature screening, one of the ways we judge whether the feature is useful is that we can calculate the IV (Information Value) to judge the importance of the feature to the result.

IV measures the amount of information of a variable. From the perspective of the formula, it is equivalent to a weighted summation of the WOE value of the independent variable. The size of the value determines the degree of influence of the independent variable on the target variable. The IV value is the difference in the

distribution of good and bad customers for a single variable. The greater the difference, the higher the discrimination between good and bad customers.

$$IV = \sum_{i=1}^{n}(\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}) * ln(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T})$$

Figure 24. Formula of IV value

where *Badi* – Bad samples and the number of group i.
Bad$_T$ – The total number of bad samples.
Goodi – Good samples and the number of group i.
Good$_T$ – The total number of good samples.

IV=sum((goodattribute-badattribute)*woe), the full name of IV is Information Value, which is generally used to compare the predictive ability of features. IV0.1 or above is considered predictive ability, and 0.2 or above is considered relatively predictive.

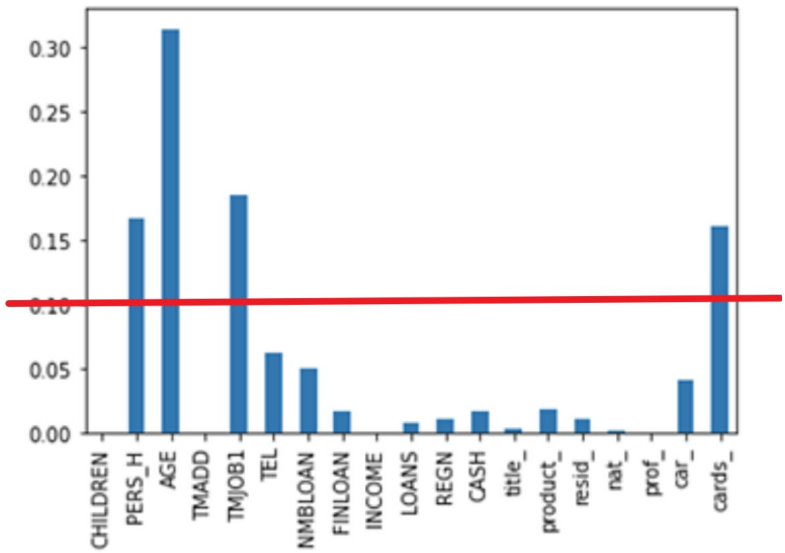


Figure 25. Plot Correlation between variables

As can be seen from the above figure, the IV values of x0, x3, x4, x7, x8, x9, x10, x11, x13, x14, x15, x16, x17 variables are significantly lower, that is, the values of CHILDREN, TMADD, TMJOB1, FINLOAN, INCOME, LOANS, REGN, CASH, title_, product_, resid_, nat_, prof_ variables. We kept variables with IV

values greater than 0.10. The IV value is significantly lower, the predictive power is poor, and its correlation with the value we want to predict GB (the dependent variable) is low, so these variables are removed.

4.4 Weight of Evidence

The Weight of Evidence (WOE) transformation can transform a logistic regression model into a standard scorecard format. Before building the model, we need to convert the filtered variables into WoE values for credit scoring.

$$WOE_i = ln(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T}) = ln(\frac{Bad_i}{Bad_T}) - ln(\frac{Good_i}{Good_T})$$

Figure 26. Formula of WOE value

where *Badi* – Bad samples and the number of group i.
BadT – The total number of bad samples.
Goodi – Good samples and the number of group i.
GoodT – The total number of good samples.

Next, woe all the required features, discard the unnecessary features, and keep only the WOE-transcoded variables:

	PERS_H_woe	AGE_woe	TMJOB1_woe	cards__woe
1371	-0.291020	-1.042	0.011	0.251162
2305	-0.291020	-0.305	-0.930	0.251162
425	0.531626	0.479	0.307	-0.638728
404	-0.349906	-0.195	-0.256	0.251162
2561	-0.327772	-0.421	0.307	0.251162
...	...	...	...	...
2134	-0.291020	-0.305	-0.005	0.251162
858	-0.349906	-1.042	0.562	0.251162
1688	0.531626	0.906	0.307	0.251162
744	0.531626	0.906	0.072	0.251162
428	0.531626	0.479	0.072	0.251162

Figure 27. WOE value of 4 variables

5.Model building

5.1 Build Logistic regression model and test accuracy (SAS)

Logistic regression is a supervised machine learning classification algorithm that is used to predict the probability of a categorical dependent variable. The dependent variable is a binary variable that contains data coded as 1 (yes/true) or 0 (no/false), used as Binary classifier (not in regression). Logistic regression can make use of large numbers of features including continuous and discrete variables and non-linear features. In Logistic Regression, the Sigmoid (aka Logistic) Function is used.

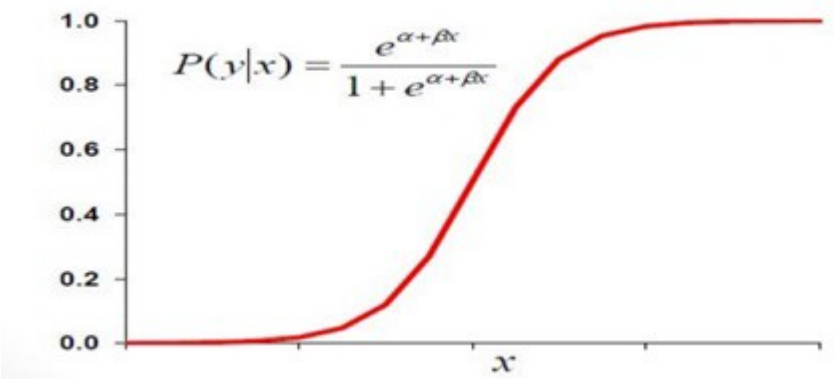


Figure 28. Sigmoid function

We want a model that predicts probabilities between 0 and 1, that is, S-shaped. There are lots of S-shaped curves. We use the logistic model: Probability = 1 / [1 +exp (B0 + b1X)] or loge[P/(1-P)] = B0 +B1X. The function on left, loge[P/(1-P)], is called the logistic function.

In any model is that we're going to split data-set into two separate sets, so i got two parts, one of them is named train, and another is named test.

Code:

```
data Work.test3_part;
set Work.test3;
```

```
label x='Random num';
x=ranuni(int(time()));
run;
data test;
 set Work.test3_part;
 if x>0.75;
run;
data train;
 set Work.test3_part;
 if x<=0.75;
run;
```

Well here it's my algorithm model. Generally we split the data-set into 75:25 ratio .what does it mean, 75 percent data take in train and 25 percent data take in test.

Code:

```
ods graphics on;
proc logistic data = train descending;
 model GB = CHILDREN PERS_H AGE INCOME TMADD title_
product_ resid_ nat_ prof_ car_ cards_TMJOB1 TEL
NMBLOAN FINLOAN EC_CARD INC INC1 BUREAU LOANS REGN
CASH/

  selection = stepwise slstay=0.15 slentry=0.15 stb;

  score data=train out = Logit_Training fitstat
outroc=troc;

  score data=test out = Logit_Validation fitstat
```

```
outroc=vroc;

Run;
```

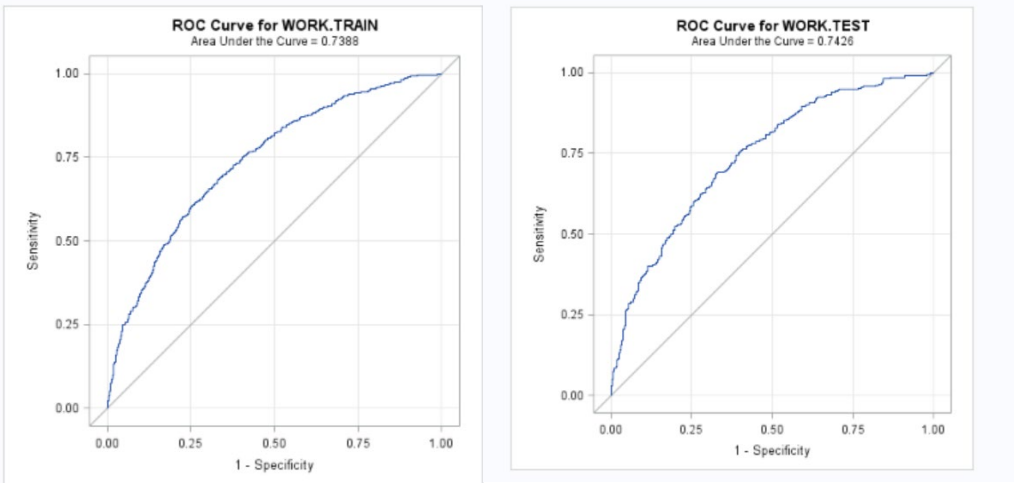


Figure 29. ROC curve for train and test

Note: No (additional) effects met the 0.15 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	AGE		1	1	132.9034		<.0001	AGE
2	cards_		1	2	56.7456		<.0001	
3	TEL		1	3	29.8345		<.0001	TEL
4	prof_		1	4	19.4814		<.0001	
5	PERS_H		1	5	16.5298		<.0001	PERS_H
6	CHILDREN		1	6	21.9025		<.0001	CHILDREN
7	LOANS		1	7	8.7263		0.0031	LOANS
8	NMBLOAN		1	8	10.1119		0.0015	NMBLOAN
9	car_		1	9	9.2467		0.0024	
10	CASH		1	10	5.6353		0.0176	CASH
11	TMJOB1		1	11	4.7610		0.0291	TMJOB1
12	nat_		1	12	3.4177		0.0645	
13	INC		1	13	2.6581		0.1030	INC
14	INCOME		1	14	4.9467		0.0261	INCOME
15	TMADD		1	15	2.5708		0.1089	TMADD

Figure 30. Summary of stepwise selection

Within SAS/STAT® 9.4. The resulting output and graphical displays are comprehensive and easily produced. We should seriously consider these valuable tools when fitting and choosing a logistic regression model. As a result of this program, a regression model was built for a test sample of data, which with curve of train = 0.7388 and curve of test = 0.7426.

5.2 Logistic regression model fitting and get ROC curve (Python)

Sklearn has the following characteristics: Simple and efficient data mining and data analysis tools. Allow everyone to reuse in complex environments. Building on NumPy, Scipy, MatPlotLib

Logistic regression is a machine learning classification algorithm used to predict the probability of categorical dependent variables. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts that P (Y = 1) is a function of X.

Logit Regression Results

Dep. Variable:	GB	**No. Observations:**	2144
Model:	Logit	**Df Residuals:**	2139
Method:	MLE	**Df Model:**	4
Date:	Thu, 09 Jun 2022	**Pseudo R-squ.:**	0.1004
Time:	18:20:05	**Log-Likelihood:**	-1336.8
converged:	True	**LL-Null:**	-1485.9
Covariance Type:	nonrobust	**LLR p-value:**	2.490e-63

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0287	0.046	-0.619	0.536	-0.120	0.062
PERS_H_woe	0.5534	0.122	4.538	0.000	0.314	0.792
AGE_woe	0.6540	0.093	7.016	0.000	0.471	0.837
TMJOB1_woe	0.6171	0.115	5.359	0.000	0.391	0.843
cards_woe	1.0315	0.117	8.786	0.000	0.801	1.262

Figure 31. Results of Logit Regression

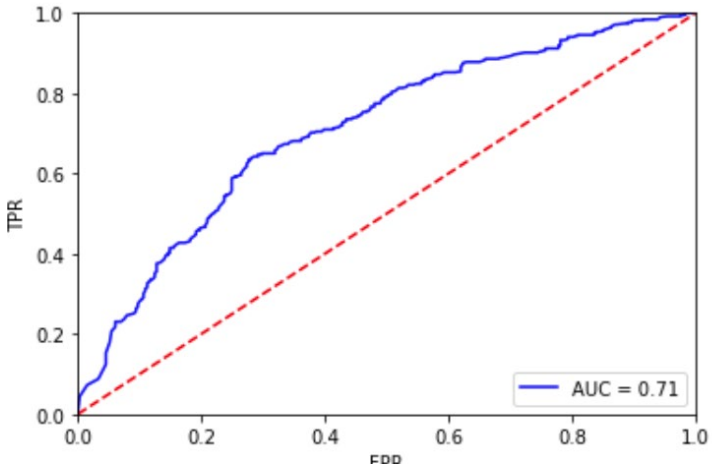


Figure 32. The ROC curve

ROC curve is another common tool used with binary classifiers. The dotted line represents the ROC curve of a pure random classifier; a good classifier is as far away from the line as possible (towards the upper left corner).

5.3 Comparison of SAS and Python

Sas: It is easy to clean data, analyze data and generate images.

Python: High development efficiency, Python has a very powerful third-party library.

In SAS, we can clearly and intuitively observe the results of data preprocessing, and it is more convenient to view the overall data set. Although our AUC in SAS is 0.738, it is better than Python's 0.71. However, Python has a wealth of various libraries. In order to facilitate subsequent calculations and the design of scorecards and GUIs, we will only use Python for subsequent processing.

5.4 Decision Tree Model Training

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

First, we have to install Graphviz drawing software:

```
[3]: pip install graphviz

Collecting graphviz
  Downloading https://files.pythonhosted.org/packages/62/dc/9dd6a6b9b897
Installing collected packages: graphviz
Successfully installed graphviz-0.14.1
Note: you may need to restart the kernel to use updated packages.
```

Figure 33. Install graphviz.

Here, we installed this version 0.14.1.

Graphviz is a visual graphics tool open sourced by AT&T Research and Lucent Bell Labs. It can be easily used to draw structured graph networks and supports multiple formats for output.

The input of Graphviz is a drawing script written in dot language. Through the analysis of the input script, the points, edges and subgraphs are analyzed, and then drawn according to the attributes.

Graphviz layout describes graphics in a simple text language, and makes charts in practical formats, such as images and SVG for web pages; PDF and Postscript that are placed in other files or displayed in an interactive graphics browser.

In the second step, we need to import the libraries..

Matplotlib is a plotting library for Python. The pyplot package encapsulates many plotting functions.

Fitting Classifier to the Training set:

```
[27]: from sklearn.tree import DecisionTreeClassifier
      classifier = DecisionTreeClassifier(criterion = 'entropy', random_state=0,class_weight={0:1,1:10})

[28]: classifier.fit(x_train,y_train)

[28]: DecisionTreeClassifier(class_weight={0: 1, 1: 10}, criterion='entropy',
                             max_depth=None, max_features=None, max_leaf_nodes=None,
                             min_impurity_decrease=0.0, min_impurity_split=None,
                             min_samples_leaf=1, min_samples_split=2,
                             min_weight_fraction_leaf=0.0, presort=False,
                             random_state=0, splitter='best')
```

Figure 34. Fitting Classifier to the Training set.

Predicting the Test set results:

```
[29]: y_pred = classifier.predict(x_test)
```

Figure 35. Predicting the Test set results.

Making the Confusion Matrix:

```
[35]: from sklearn.metrics import confusion_matrix
      cm = confusion_matrix(y_test, y_pred)

      scoretest = classifier.score(x_test,y_test)
```

Figure 36. Making the Confusion Matrix.

Visual decision tree:

```
[30]: from IPython.display import Image
      import pydotplus
      from sklearn import tree
```

```
[31]: dot_data = tree.export_graphviz(classifier, out_file=None,
                                     filled=True, rounded=True,
                                     special_characters=True)
      graph = pydotplus.graph_from_dot_data(dot_data)
```

Figure 37. Visual decision tree.

We save the output picture to the desktop named ‘out’:

```
[32]: graph.write_png("C:/Users/Administrator/Desktop/out.png")

[32]: True
```

Figure 38. Output the picture.

If the result is True, the code runs successfully.

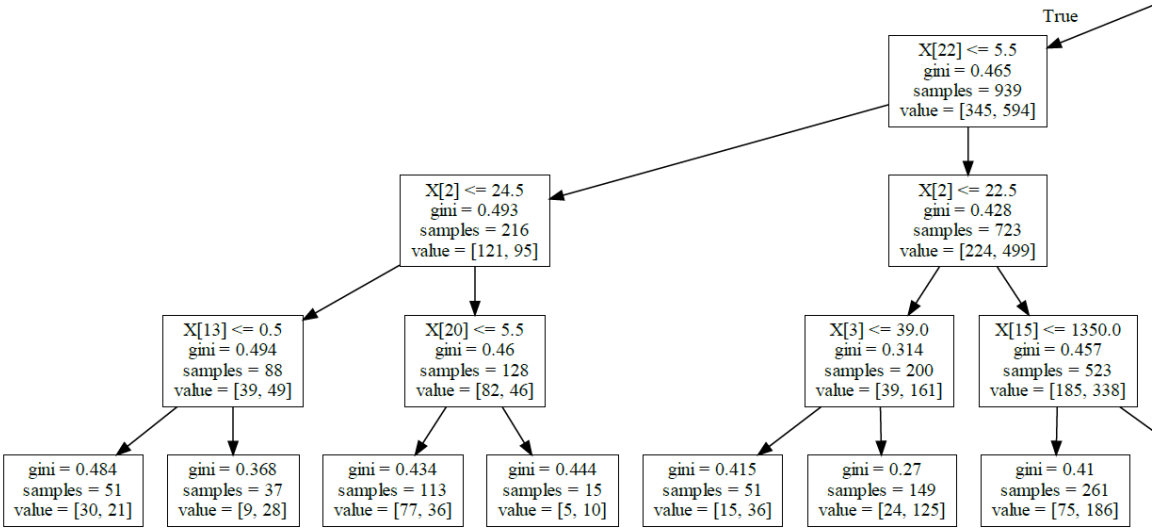


Figure 39. Decision tree.

5.5 Random forests Model Training

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

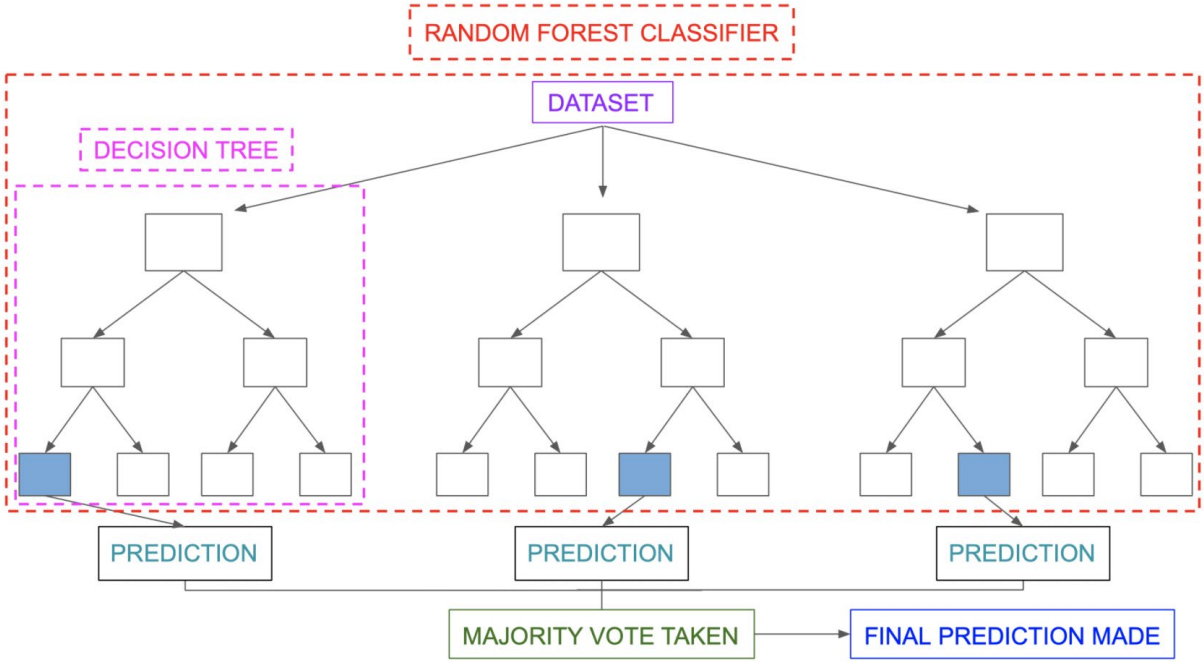


Figure 40. Model of Random Forest

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

```
7]: # Use the model to predict results
    model.fit(X_trian, y_trian)

7]: RandomForestClassifier(max_depth=5, min_samples_leaf=10, n_estimators=12,
                           random_state=2022)

8]: y_pred = model.predict(X_test)

9]: acc = round(accuracy_score(y_test, y_pred), 3)
    print('Accuracy:', acc)
    # Precision
    pre = round(precision_score(y_test, y_pred), 3)
    print('Precision:',pre)
    # recall
    recall = round(recall_score(y_test, y_pred), 3)
    print('recall:',recall)
    # f1-score
    f1 = round(f1_score(y_test, y_pred), 3)
    print('f1-scoreA:',f1)

    Accuracy: 0.677
    Precision: 0.688
    recall: 0.677
    f1-scoreA: 0.682
```

Figure 41. Fitting model.

Parameter optimization:

```
# parameter optimization!
parameters = {'n_estimators':[9,10,11,12,13], 'max_depth':[3, 4, 5, 6, 7],'min_samples_leaf':[14,15,16,17]}
new_model = RandomForestClassifier(random_state=2022)
grid_search = GridSearchCV(new_model, parameters, cv=6, scoring='accuracy')
grid_search.fit(X_trian, y_trian)
print(grid_search.best_params_)
```

```
{'max_depth': 6, 'min_samples_leaf': 16, 'n_estimators': 12}
```

```
model = RandomForestClassifier(max_depth=6, n_estimators=12, min_samples_leaf=16, random_state=2022)
```

```
model.fit(X_trian, y_trian)
```

```
RandomForestClassifier(max_depth=6, min_samples_leaf=16, n_estimators=12,
                       random_state=2022)
```

Figure 42. Parameter optimization.

Estimate the importance of features:

```
# Estimate the importance of features
features = X.columns
importances = model.feature_importances_
b = pd.DataFrame()
b['feature'] = features
b['importance'] = importances
b = b.sort_values('importance', ascending=False)
print(b)
```

```
  feature  importance
1     AGE    0.438885
2  TMJOB1    0.224006
3  cards_    0.178142
0  PERS_H    0.158966
```

Figure 43. Estimate the importance of features.

Get classification performance metrics:

```
# Get classification performance metrics
# Accuracy
acc = round(accuracy_score(y_test, y_pred), 3)
print('Accuracy:', acc)
# Precision
pre = round(precision_score(y_test, y_pred), 3)
print('Precision:',pre)
# recall
recall = round(recall_score(y_test, y_pred), 3)
print('recall:',recall)
# f1-score
f1 = round(f1_score(y_test, y_pred), 3)
print('f1-scoreA:',f1)
```

```
Accuracy: 0.661
Precision: 0.664
recall: 0.685
f1-scoreA: 0.674
```

Figure 44. Get classification performance metrics.

We save the output picture to the desktop:

```
[81]: import os
      from sklearn.datasets import load_iris
      from sklearn.tree import export_graphviz
      import graphviz

[82]: for idx, estimator in enumerate(model.estimators_):
          dot_data = export_graphviz(estimator, out_file=None)
          graph = graphviz.Source(dot_data)
          graph.render(f'Decision tree{idx + 1} of Random Forest')
```

Figure 45. Save the picture

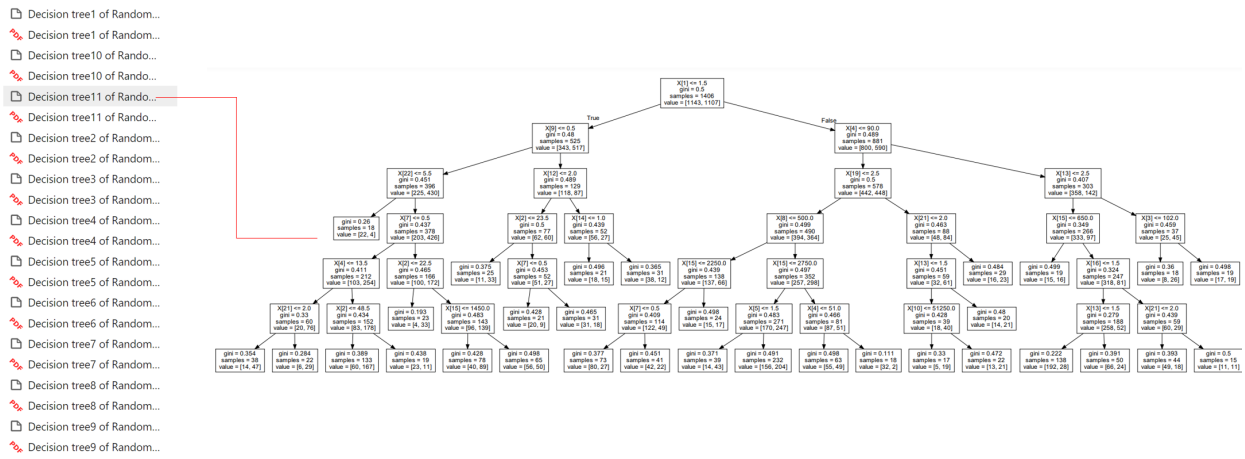


Figure 46. Visual one of the Random Forest

6. Compare the Random forests to other models

Any algorithm has limitations, so there is no "universal optimal algorithm", only a certain algorithm may be asymptotically optimal in a specific situation. Therefore, it is very important to evaluate the algorithm performance and choose the optimal algorithm.

First of all, what we often say is to choose a correct evaluation standard. Common ones are: accuracy, recall, precision, ROC, Precision-Recall Curve, F1, etc.

Here I choose ROC as the evaluation criterion:

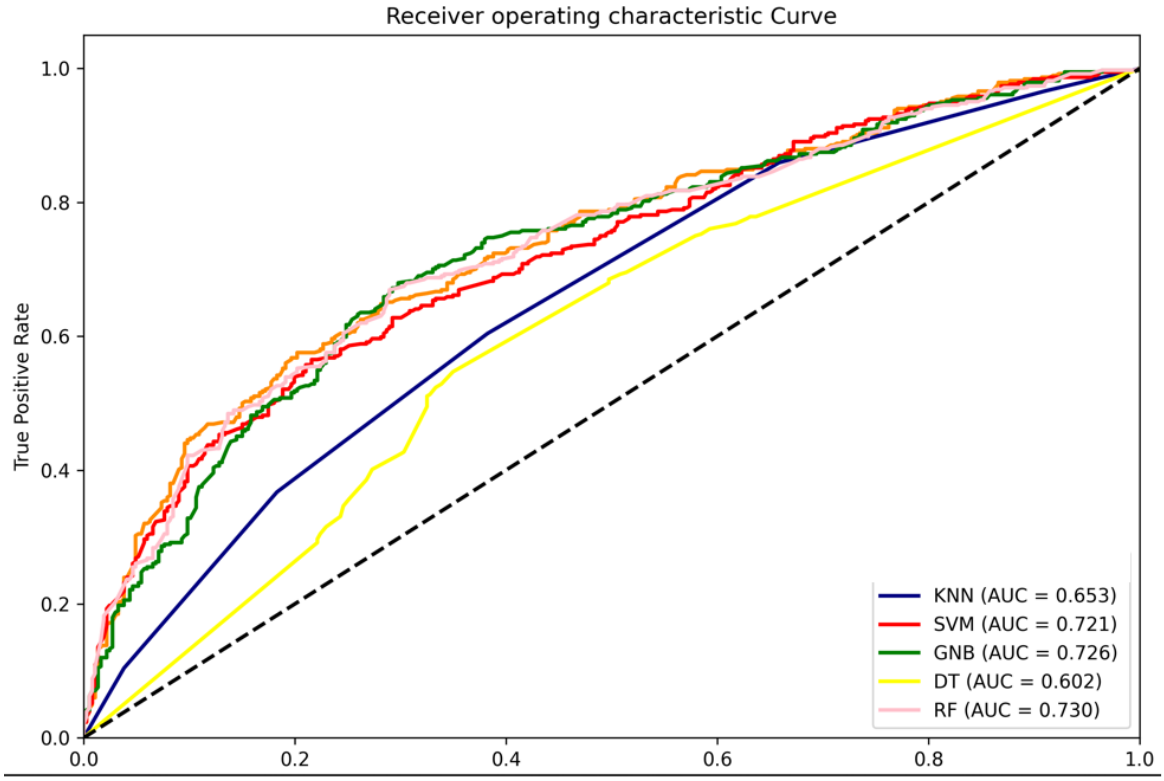


Figure 47. Compare the Random forests to other models

By comparing the AUC values of various models, we find that the accuracy of Random forests 73.0% significantly better than the accuracy of Other models. So, we finally choose the random forest model to find the right customers to help banks control loan risk.

Code:

```
from sklearn.metrics import confusion_matrix, accuracy_score, f1_score, roc_auc_score, recall_score, precision_score, \
    roc_curve
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import KFold
```

```
  from sklearn.discriminant_analysis import
LinearDiscriminantAnalysis
  from sklearn.linear_model import LogisticRegression
  from sklearn.svm import SVC
  from sklearn.naive_bayes import GaussianNB
  from sklearn.model_selection import cross_val_score
  from matplotlib import pyplot
  from sklearn.neighbors import KNeighborsClassifier
  from sklearn.model_selection import train_test_split
  from sklearn.metrics import classification_report
  from sklearn.ensemble import RandomForestClassifier
  from xgboost import XGBClassifier
  from sklearn.model_selection import cross_validate
  from sklearn.metrics import roc_curve, auc
  import matplotlib.pyplot as plt
  import pandas as pd
  import numpy as np


  df =
pd.read_csv(r'C:\Users\Administrator\Desktop\scientific
_work\2.csv')
  df = df[df['PERS_H']<10]
  X = df[['PERS_H', 'AGE', 'TEL', 'NMBLOAN', 'car_',
'cards_']]
```

```
  y = df['GB']
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=2022)


  def calculate_auc(y_test, pred):
      print("auc:", roc_auc_score(y_test, pred))
      fpr, tpr, thersholds = roc_curve(y_test, pred)
      roc_auc = auc(fpr, tpr)
      plt.plot(fpr, tpr, 'k-', label='ROC (area = {0:.2f})'.format(roc_auc), color='blue', lw=2)
      plt.xlim([-0.05, 1.05])
      plt.ylim([-0.05, 1.05])
      plt.xlabel('False Positive Rate')
      plt.ylabel('True Positive Rate')
      plt.title('ROC Curve')
      plt.legend(loc="lower right")
      plt.plot([0, 1], [0, 1], 'k--')
      plt.show()


  def Find_Optimal_Cutoff(TPR, FPR, threshold):
      y = TPR - FPR
      Youden_index = np.argmax(y)   # Only the first occurrence is returned.
```

```
    optimal_threshold = threshold[Youden_index]
    point = [FPR[Youden_index], TPR[Youden_index]]
    return optimal_threshold, point


def ROC(label, y_prob):
    fpr, tpr, thresholds = roc_curve(label, y_prob)
    roc_auc = auc(fpr, tpr)
    optimal_threshold, optimal_point = Find_Optimal_Cutoff(TPR=tpr, FPR=fpr, threshold=thresholds)
    return fpr, tpr, roc_auc, optimal_threshold, optimal_point


def calculate_metric(label, y_prob, optimal_threshold):
    p = []
    for i in y_prob:
        if i >= optimal_threshold:
            p.append(1)
        else:
            p.append(0)
    confusion = confusion_matrix(label, p)
    print(confusion)
    TP = confusion[1, 1]
```

```
        TN = confusion[0, 0]
        FP = confusion[0, 1]
        FN = confusion[1, 0]
        Accuracy = (TP + TN) / float(TP + TN + FP + FN)
        Sensitivity = TP / float(TP + FN)
        Specificity = TN / float(TN + FP)
        return Accuracy, Sensitivity, Specificity


  models = [('Logit', LogisticRegression(max_iter=5000)),
            ('KNN', KNeighborsClassifier()),
            ('SVM', SVC(probability=True)),
            ('GNB', GaussianNB()),
            ('DT',
DecisionTreeClassifier(random_state=0)),
            ('RF',   RandomForestClassifier(max_depth=2,
random_state=0))]


  results = []
  roc_ = []


  for name, model in models:
        clf = model.fit(X_train, y_train)
        pred_proba = clf.predict_proba(X_test)
```

```
      y_prob = pred_proba[:, 1]

      fpr, tpr, roc_auc, Optimal_threshold, optimal_point
= ROC(y_test, y_prob)


      Accuracy,      Sensitivity,      Specificity      =
calculate_metric(y_test, y_prob, Optimal_threshold)

      result = [Optimal_threshold, Accuracy, Sensitivity,
Specificity, roc_auc, name]

      results.append(result)

      roc_.append([fpr, tpr, roc_auc, name])


  df_result = pd.DataFrame(results)

  df_result.columns = ["Optimal_threshold", "Accuracy",
"Sensitivity", "Specificity", "AUC_ROC", "Model_name"]


  color = ["darkorange", "navy", "red", "green", "yellow",
"pink"]

  plt.figure()

  plt.figure(figsize=(10, 10))

  lw = 2

  plt.plot(roc_[0][0], roc_[0][1], color=color[0], lw=lw,
label=roc_[0][3] + ' (AUC = %0.3f)' % roc_[0][2])

  plt.plot(roc_[1][0], roc_[1][1], color=color[1], lw=lw,
label=roc_[1][3] + ' (AUC = %0.3f)' % roc_[1][2])
```

```
  plt.plot(roc_[2][0], roc_[2][1], color=color[2], lw=lw,
label=roc_[2][3] + ' (AUC = %0.3f)' % roc_[2][2])

  plt.plot(roc_[3][0], roc_[3][1], color=color[3], lw=lw,
label=roc_[3][3] + ' (AUC = %0.3f)' % roc_[3][2])

  plt.plot(roc_[4][0], roc_[4][1], color=color[4], lw=lw,
label=roc_[4][3] + ' (AUC = %0.3f)' % roc_[4][2])

  plt.plot(roc_[5][0], roc_[5][1], color=color[5], lw=lw,
label=roc_[5][3] + ' (AUC = %0.3f)' % roc_[5][2])

  plt.plot([0,  1],  [0,  1],  color='black',  lw=lw,
linestyle='--')

  plt.xlim([0.0, 1.0])

  plt.ylim([0.0, 1.05])

  plt.xlabel('False Positive Rate')

  plt.ylabel('True Positive Rate')

  plt.title('Receiver operating characteristic Curve')

  plt.legend(loc="lower right")

  plt.savefig("roc_curve.png", dpi=300)

  plt.show()
```

7. Model checking and building scorecards

After the model is trained, we need to convert each bin (that is, the value segment) for each variable into a specific score.

The scorecard does not directly use the customer default rate p, but uses the ratio of the default probability to the normal probability, called Odds, that is, the scorecard will map odds to score. Odds means probability, probability, refers to the ratio of the probability (probability) of an event to the probability (probability) of

not occurring.

$$odds1=p1/(1-p1)=A/B$$

Among them, A and B are constants, and the reason for the negative sign in front of B is that the lower the default probability, the higher the score. Because in actual business, the higher the score, the lower the risk.\

Benchmark score. The benchmark score θ is the score at a certain ratio. Here we set it to 112. The reason for setting it to 112 is so that the value of q, which is our base Score, is exactly 100.

```
p = 5/np.log(2)
q = 112 - 5*np.log(5)/np.log(2)
```

Figure 48. Score calculation

PDO (point of double), the change in the score when the ratio doubles. Suppose we set the score to decrease by 5 when odds double.

Table 2. Feature Score Table

Variable	Interval	Score
base	**-**	**100**
PERS_H	<=1	2.0
	(1,2]	-1.0
	(2,3]	-1.0
	(3,5]	-1.0
	>=5	0.0
AGE	<=23.0	4.0
	(23.0, 25.0]	2.0
	(25.0, 28.0]	1.0
	(28.0, 31.0]	0.0
	(31.0, 35.0]	-1.0
	(35.0, 39.0]	-1.0
	(39.0, 46.0]	-2.0
	>=46.0	-5.0
TMJOB1	<=12.0	3.0
	(12.0, 21.0]	1.0

	(21.0, 33.0]	0.0
	(33.0, 45.0]	0.0
	(45.0, 72.0]	-0.0
	(72.0, 144.0]	-1.0
	>=144.0	-4.0
cards__	<=1.0	-0.0
	(1.0, 2.0]	-5.0
	(2.0, 5.0]	-6.0
	>=5.0	2.0

Import our test data:

	PERS_H	AGE	TMJOB1	cards_	GB	Score
1371	3	47	42	7	1	95.0
2305	3	37	216	7	0	95.0
425	1	25	15	2	1	94.0
404	2	34	96	7	1	98.0
2561	4	40	15	7	0	100.0
1402	2	29	18	7	1	101.0
50	1	27	15	2	1	92.0
1228	4	32	9	7	1	103.0
2571	4	36	240	7	0	96.0
2044	2	58	36	2	0	82.0
1246	3	35	12	7	1	102.0
1054	4	41	39	7	1	99.0
1791	1	25	84	2	0	92.0
470	2	23	24	7	1	106.0
77	1	30	15	7	1	104.0
446	3	34	9	7	1	102.0
1616	3	37	240	7	0	95.0
975	3	21	15	7	1	105.0
2393	3	31	144	2	0	81.0
1475	1	66	72	3	1	89.0

Figure 49. Scoring Tests and Results

8. GUI scorecard

It is convenient for bank managers to operate. We need to create a simple GUI visualization program, enter part of the user's information, and then get the customer's rating value, which will help the bank to evaluate whether it can get a loan.

Figure 50. Enter test client information

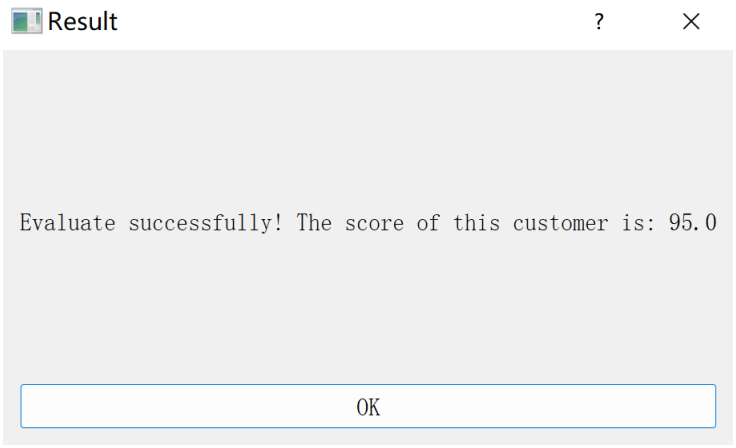


Figure 51. Get the rating results

Conclusion

In this project, we preprocessed the banking data using Python and manually cleaned the Outliers based on boxplots. For the preprocessing of categorical variables, we have made a detailed introduction. Feature selection was performed by checking the correlation of variables and Information Value, and finally 4 important variables were screened out.

By comparing the AUC values of various models, we find that the accuracy of Random forests 73.0% significantly better than the accuracy of Other models. So, we finally choose the random forest model to find the right customers to help banks control loan risk.

Finally, we use PyQT5 to make a visual scorecard, enter the user's four information, we can get the user's accurate score, help the bank to judge the value of the customer

9. Financial management, resource efficiency and resource saving

The purpose of this section discusses the issues of competitiveness, resource efficiency and resource saving, as well as financial costs regarding the object of study of Master's thesis. Competitiveness analysis is carried out for this purpose. SWOT analysis helps to identify strengths, weaknesses, opportunities and threats associated with the project, and give an idea of working with them in each particular case. For the development of the project requires funds that go to the salaries of project participants and the necessary equipment, a complete list is given in the relevant section. The calculation of the resource efficiency indicator helps to make a final assessment of the technical decision on individual criteria and in general.

9.1 Competitiveness analysis of technical solutions

In order to find sources of financing for the project, it is necessary, first, to determine the commercial value of the work. Analysis of competitive technical solutions in terms of resource efficiency and resource saving allows to evaluate the comparative effectiveness of scientific development. This analysis is advisable to carry out using an evaluation card.

First of all, it is necessary to analyze possible technical solutions and choose the best one based on the considered technical and economic criteria.

Evaluation map analysis presented in Table 1. The position of your research and competitors is evaluated for each indicator by you on a five-point scale, where 1 is the weakest position and 5 is the strongest. The weights of indicators determined by you in the amount should be 1. Analysis of competitive technical solutions is determined by the formula:

$$C = \sum W_i \cdot P_i,,$$

C - the competitiveness of research or a competitor;

Wi– criterion weight;

Pi – point of i-th criteria.

P_{i1} – Accuracy of the obtained dataset

P_{i2} – Scoring Metrics for Prediction Accuracy

Many banks have begun to try to use big data to drive business operations. For example, the China CITIC Bank Credit Card Center uses big data technology to achieve real-time marketing, China Everbright Bank has established a social network information database, and China Merchants Bank uses big data to develop small and micro loans.

Our project helps bank managers judge whether a new user is worthy of a loan by evaluating the presentation of manufactured cards and visualizing the results. Models from other banks are currently slightly lacking in visual representation, but our visual display interface is concise and clear.

Table 3. Evaluation card for comparison of competitive technical solutions

Evaluation criteria	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Ease of operation	0.10	5	3	5	0.50	0.30	0.50
2. Enough data	0.30	4	4	5	1.2	1.2	1.5
3. Smart interface quality	0.05	3	5	5	0.15	0.25	0.25
4. Ability to connect to PC	0.05	4	4	4	0.20	0.20	0.20
Economic criteria for performance evaluation							
1. Competitive power	0.10	4	3	5	0.4	0.3	0.5
2. Prediction accuracy	0.30	5	4	4	1.5	1.2	1.2
3. Development cost	0.05	3	3	3	0.15	0.15	0.15
4. After-sale service	0.05	4	5	4	0.2	0.25	0.2

Total	**1**	**32**	**31**	**35**	**4.3**	**3.85**	**4.5**

At present, the commercial application of big data in banks is mainly based on their own transaction data and customer data, external data is supplemented by descriptive data analysis, predictive data modeling is supplemented, and business customers are the main business. Operating products as a supplement.

9.2 SWOT analysis

Complex analysis solution with the greatest competitiveness is carried out with the method of the SWOT analysis: Strengths, Weaknesses, Opportunities and Threats. The analysis has several stages. The first stage consists of describing the strengths and weaknesses of the project, identifying opportunities and threats to the project that have emerged or may appear in its external environment. The second stage consists of identifying the compatibility of the strengths and weaknesses of the project with the external environmental conditions. This compatibility or incompatibility should help to identify what strategic changes are needed.

Table 4. SWOT analysis

	Strengths: S1. Some of the information with commercial value can be analyzed through data mining. S2. Identify fraudulent transactions and help banks reduce risk. S3. Big data analysis can improve business decisions and provide data support.	**Weaknesses:** W1. Noisy data in the dataset needs to be preprocessed. W2. Requires analysis using a large dataset to improve accuracy. W3. Binding to the Data, a dataset only applies to one bank.
Opportunities:	*Strategy which based on*	*Strategy which based on*

O1. Conduct risk assessments on lenders, identify fraudulent transactions, and help bank managers reduce risk. O2. Visual processing of customer information, intuitive display of various scoring indicators, convenient for managers to operate.	*strengths and opportunities: Analyze the score of borrowers through machine learning methods, and visualize the information.*	*weaknesses and opportunities: Use different data preprocessing methods whenever possible and analyze the results multiple times.*
Threats: T1. The prediction accuracy is not ideal. T2. The data set prepared by the bank is not large enough, which leads to the problem of prediction overfitting.	*Strategy which based on strengths and threats: The initial data needs to be feature-engineered before prediction.*	*Strategy which based on weaknesses and threats: Use cross-validation methods to prevent overfitting problems when performing data analysis.*

9.3 Project Initiation

The initiation process group consists of processes that are performed to define a new project or a new phase of an existing one. In the initiation processes, the initial purpose and content are determined and the initial financial resources are fixed. The internal and external stakeholders of the project who will interact and influence the overall result of the research project are determined.

Table 5. Stakeholders of the project

Project stakeholders	**Stakeholder expectations**
Bank	Easy to use, high accuracy of scoring model.
Customers of bank	Understanding the Bank's Scoring System.

Table 6. Purpose and results of the project

Purpose of project:	This project aims to screen credit customers according

	to the personal information on the finance dataset, finding the right person (the good customer).
Expected results of the project:	Display various scoring indicators in a visual way, and analyze customers who can be credited.
Criteria for acceptance of the project result:	The default rate of bank users who are guaranteed to pass the bank loan review is 10% lower than expected.
Requirements for the project result:	1. The project must be completed before May 31, 2022 of the current year.
	2. The results obtained must meet the acceptance criteria of the project results.

It is necessary to solve the some questions: who will be part of the working group of this project, determine the role of each participant in this project, and prescribe the functions of the participants and their number of labor hours in the project.

Table 7. Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours (working days (from table 7) × 6 hours)
1	Supervisor	Head of project	Suggest project direction and review master's thesis.	384 hours
2	Student	Executor	1. Analyze datasets find good customers. 2.Writing master's dissertations.	816 hours

Project limitations are all factors that can be as a restriction on the degree of freedom of the project team members.

Table 8. Project limitations

Factors	**Limitations / Assumptions**
3.1. Project's budget	525000 RUB
3.1.1. Source of financing	TPU
3.2. Project timeline:	5/10/2020 to 30/05/2022
3.2.1. Date of approval of plan of project	15/11/2020
3.2.2. Completion date	18/05/2022

As part of planning a science project, you need to build a project timeline and a Gantt Chart.

Table 9. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
General Technical supervision	22 days	5/10/2020	5/11/2020	Supervisor
Planning project	21 days	5/11/2020	5/12/2020	Supervisor
Data preprocessing and ROC curve plotting	15 days	5/12/2020	5/01/2021	Student
Plot the logistic regression curve and build the model	20 days	1/03/2021	1/04/2021	Student
Building decision trees and random forest models	39 days	1/05/2021	29/06/2021	Student
Compare the classification accuracy of decision trees and random forests	21 days	15/10/2021	15/11/2021	Supervisor/ Student

Visualize scoring metrics using GUI	21 days	06/03/2022	06/04/2022	Student
Preparing of dissertation	20 days	30/04/2022	30/05/2022	Student

A Gantt chart, or harmonogram, is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.

Table 10. A Gantt chart

№	Activities	Participants	T_c, days	Duration of the project											
				2020			2021						2022		
				10	11	12	1	3	5	7	9	11	3	4	5
1	General Technical supervision	Supervisor	22												
2	Planning project	Supervisor	21												
3	Data preprocessing and ROC curve plotting	Student	15												
4	Plot the logistic regression curve and build the model	Student	20												
5	Building decision trees and random forest models	Student	39												
6	Compare the classification accuracy of decision trees and random forests	Supervisor/ Student	21												

7	Visualize scoring metrics using GUI	Student	21												
8	Preparing of dissertation	Student	20												

9.4 Scientific and technical research budget

The amount of costs associated with the implementation of this work is the basis for the formation of the project budget. This budget will be presented as the lower limit of project costs when forming a contract with the customer.

To form the final cost value, all calculated costs for individual items related to the manager and the student are summed.

In the process of budgeting, the following grouping of costs by items is used:

Material costs of scientific and technical research;

costs of special equipment for scientific work (Depreciation of equipment used for design);

basic salary;

additional salary;

labor tax;

overhead.

Calculation of material costs

The calculation of material costs is carried out according to the formula:

$$C_m = (1 + k_T) \cdot \sum_{i=1}^{m} P_i \cdot N_{consi}$$

where m – the number of types of material resources consumed in the performance of scientific research;

N_{consi} – the amount of material resources of the i-th species planned to be used when performing scientific research (units, kg, m, m^2, etc.);

P_i – the acquisition price of a unit of the i-th type of material resources consumed

(rub./units, rub./kg, rub./m, rub./m^2, etc.);

k_T – coefficient taking into account transportation costs.

Prices for material resources can be set according to data posted on relevant websites on the Internet by manufacturers (or supplier organizations).

Table 11. Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Electricity of computer	kWh	150	5.8	870
A4 Papers		100	1.0	100
Printing A4 Papers		10	80	800
Total				1770

9.5 Costs of special equipment

This point includes the costs associated with the acquirement of special equipment (instruments, stands, devices and mechanisms) necessary to carry out work on a specific topic.

Table 12. Costs of special equipment (+software)

№	equipment identification	Quantity of equipment	Price per unit, rub.	Total cost of equipment, rub.
1.	Rog computer	1	120000	120000
2.	Pycharm software	1	12010	12010
Total		132010		

9.6 Basic salary

This point includes the basic salary of participants directly involved in the implementation of work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system The basic salary (S_b) is calculated according to the formula:

$$S_b = S_a \cdot T_w ,$$

where S_b – basic salary per participant;

T_w – the duration of the work performed by the scientific and technical worker, working days;

S_d - the average daily salary of an participant, rub.

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v} ,$$

где S_m – monthly salary of an participant, rub .;

M – the number of months of work without leave during the year:

at holiday in 48 days, M = 11.2 months, 6 day per week;

F_v – valid annual fund of working time of scientific and technical personnel (251 days).

Table 13. The valid annual fund of working time

Working time indicators	
Calendar number of days	365
The number of non-working days - weekend - holidays	 52 14
Loss of working time - vacation - isolation period - sick absence	 48
The valid annual fund of working time	251

Monthly salary is calculated by formula:

$$S_{month} = S_{base} \cdot (k_{premium} + k_{bonus}) \cdot k_{reg} , \quad (x)$$

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 14. Calculation of the base salaries

Performers	S_{base}, rubles	$k_{premium}$	k_{bonus}	k_{reg}	S_{month}, rub.	W_d, rub.	T_p, work days	W_{base}, rub.
Supervisor	37700			1.3	49010	1633.7	64	104554.7
Student	19200				24960	832	136	113152

Additional salary

This point includes the amount of payments stipulated by the legislation on labor, for example, payment of regular and additional holidays; payment of time associated with state and public duties; payment for work experience, etc. Additional salaries are calculated on the basis of 10-15% of the base salary of workers:

$$W_{add} = k_{extra} \cdot W_{base},$$

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient (10%);

W_{base} – base salary, rubles.

Labor tax

Tax to extra-budgetary funds are compulsory according to the norms established by the legislation of the Russian Federation to the state social insurance (SIF), pension fund (PF) and medical insurance (FCMIF) from the costs of workers.

Payment to extra-budgetary funds is determined of the formula:

$$P_{social} = k_b \cdot (W_{base} + W_{add})$$

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conducting educational and

scientific activities have rate - 27.1%.

Table 15. Labor tax

	Project leader	**Engineer**
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	104554.7	113152
Labor tax, rubles	28334.32	30664.19

Overhead costs

Overhead costs include other management and maintenance costs that can be allocated directly to the project. In addition, this includes expenses for the maintenance, operation and repair of equipment, production tools and equipment, buildings, structures, etc.

Overhead costs account from 30% to 90% of the amount of base and additional salary of employees.

Overhead is calculated according to the formula:

$$C_{ov} = k_{ov} \cdot (W_{base} + W_{add})$$

where k_{ov} – overhead rate.

Table 16. Overhead

	Project leader	**Engineer**
Overhead rate	30%	
Salary, rubles	104554.7	113152
Overhead, rubles	31366.41	33945.6

Formation of budget costs

The calculated cost of research is the basis for budgeting project costs. Determining the budget for the scientific research is given in the table 16.

Table 17. Items expenses grouping

Name	Cost, rubles
Material costs	1770
Equipment costs	132010
Basic salary	217706.7
Additional salary	0
Labor tax	58008.51
Overhead	65312.01
Other direct costs	0
Total planned costs	**474807.2**

9.7 Evaluation of the comparative effectiveness of the project

Determination of efficiency is based on the calculation of the integral indicator of the effectiveness of scientific research. Its finding is associated with the definition of two weighted average values: financial efficiency and resource efficiency.

The integral indicator of the financial efficiency of a scientific study is obtained in the course of estimating the budget for the costs of three (or more) variants of the execution of a scientific study. For this, the largest integral indicator of the implementation of the technical problem is taken as the calculation base (as the denominator), with which the financial values for all the options are correlated.

The integral financial measure of development is defined as:

,

where – integral financial measure of development;

C_i – the cost of the i-th version;

C_{max} – the maximum cost of execution of a research project (including analogues).

The obtained value of the integral financial measure of development reflects the corresponding numerical increase in the budget of development costs in times (the

value is greater than one), or the corresponding numerical reduction in the cost of development in times (the value is less than one, but greater than zero).

Since the development has one performance, then = 1.

The integral indicator of the resource efficiency of the variants of the research object can be determined as follows:

$$I_m^a = \sum_{i=1}^{n} a_i b_i^a \qquad I_m^p = \sum_{i=1}^{n} a_i b_i^p$$

where I_m – integral indicator of resource efficiency for the i-th version of the development;

a_i – the weighting factor of the i-th version of the development;

b_i^a, b_i^p – score rating of the i-th version of the development, is established by an expert on the selected rating scale;

n – number of comparison parameters.

The calculation of the integral indicator of resource efficiency is presented in the form of table 18.

Table 18 – Evaluation of the performance of the project

Criteria / **Weight criterion**	**Weight criterion**	**Points**
1. Ease of operation	0.10	13
2. Enough data	0.30	12
4. Smart interface quality	0.05	13
5. Ability to connect to PC	0.05	12
Economic criteria for performance evaluation		
1. Competitive power	0.10	12
2. Prediction accuracy	0.30	13
3. Development cost	0.05	9
4. After-sale service	0.05	13

Total	1	97

The integral indicator of the development efficiency ( ) is determined on the basis of the integral indicator of resource efficiency and the integral financial indicator using the formula:

, $I_{\text{исп.2}} = \frac{I_{\text{р-исп2}}}{I_{\text{финр}}^{\text{исп.2}}}$ и т.д.

Comparison of the integral indicator of the current project efficiency and analogues will determine the comparative efficiency. Comparative effectiveness of the project:

.

Thus, the effectiveness of the development is presented in table 19.

Table 19 – Efficiency of development

№	**Indicators**	**Points**
1	Integral financial measure of development	12
2	Integral indicator of resource efficiency of development	15
3	Integral indicator of the development efficiency	13

Comparison of the values of integral performance indicators allows us to understand and choose a more effective solution to the technical problem from the standpoint of financial and resource efficiency.

9.8 Conclusion for Financial management, resource efficiency and resource saving

Thus, in this section was developed stages for design and create competitive development that meet the requirements in the field of resource efficiency and resource saving.

These stages includes:

development of a common economic project idea, formation of a project concept;

organization of work on a research project;

identification of possible research alternatives;

research planning;

assessing the commercial potential and prospects of scientific research from the standpoint of resource efficiency and resource saving;

determination of resource (resource saving), financial, budget, social and economic efficiency of the project.

10. Social responsibility

10.1 Introduction

The developed project aims to screen credit customers according to the personal information on the finance dataset, finding the right person (the good customer) to provide a reference for the bank's loan business. The development of the program is only carried out with the help of computer.

In this section, harmful and dangerous factors affecting the work of personnel will be considered, the impact of the developed program on the environment, legal and organizational issues, measures in emergency situations will be considered.

The work was carried out in the hall of residence of TPU (2th floor) and my apartment in China. Room 402 was a research execution place. The layout of the apartment is shown in Figure:

Figure 52. Apartment layout 402

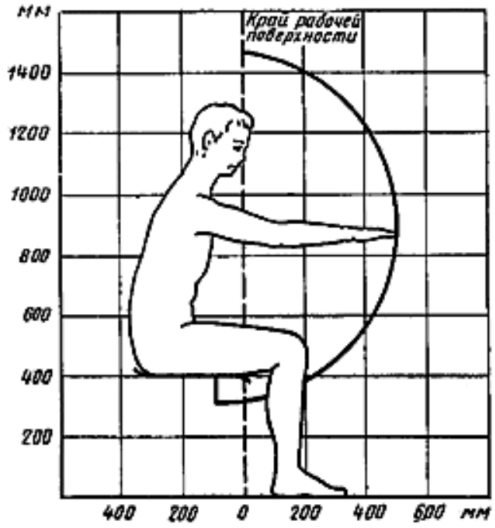
10.2 Legal and organizational issues of occupational safety

Today, one of the main ways to fundamentally improve all prevention efforts is the widespread implementation of an integrated occupational safety and health management system to reduce overall accident rates and occupational morbidity. This means combining isolated activities into a targeted system of action at all levels and stages of the production process.

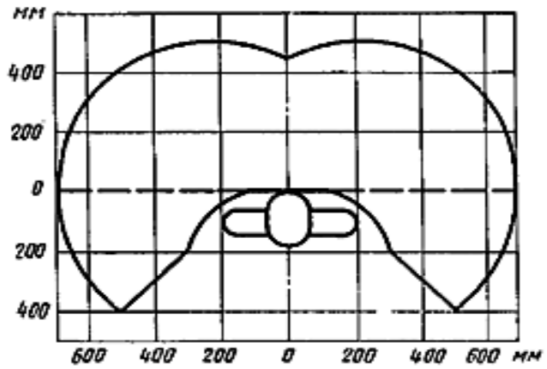
Occupational safety is a multidisciplinary field concerned with the safety, health, and welfare of people at work. The goal of occupational safety programs is to foster a safe and healthy work environment. Occupational safety may also protect co-workers, family members, employers, customers, and many others who might be affected by the workplace environment.

According to the GOST 12.2.032-78 SSBT - Workplace when performing work while sitting General ergonomic requirements:

- The design of the workplace and the relative position of all its elements (seats, controls, ways of displaying information, etc.) must conform to anthropometric, physiological and psychological requirements and the nature of the work;
- The shape of the work surface for each type of equipment should be determined according to the nature of the work performed. It can be rectangular, have a cutout for a work box or a groove for a desktop, etc. If necessary, a handrail should be installed on the work surface;
- The height of the footrest must be adjustable. The width must be at least 300mm and the length must be at least 400mm. The bracket surface must have grooves. A 10mm high edge should be provided along the front edge;
- The design of the workplace should ensure that labor is carried out within the confines of the sports field. Figures 1 and 2 show the extent of a sports field of average human size in both vertical and horizontal planes.(Reach zone of the motor field in the horizontal plane with a working surface height above the floor of 725 mm)



Figures 53. Range of reach of the motor field in the vertical plane



Figures 54. Reach zone of the motor field in the horizontal plane

- The frequently used way of displaying information requires less accurate and faster readings, which can be placed in the vertical plane at an angle of ±30° from the normal line of sight, and in the horizontal plane at an angle of ±30° from the sagittal plane;

According to the SanPiN 2.2.2/2.4.1340-03, workplaces where PCs are used must comply with these hygiene rules:

- In industrial sites where computer-assisted work is used, the temperature, relative humidity and air velocity of the workplace must comply with the current hygiene standards for the microclimate of the industrial site.
- In PC-equipped locations, routine wet cleaning and system ventilation are performed after every hour of work on the PC.
- The content of positive and negative ions in the air of the place where the PC is

located must meet the current hygiene and epidemiological standards.
- The table should be positioned so that the video display terminal faces the side of the light opening so that natural light falls mainly on the left side.
- Workplaces using PCs are recommended to be separated from each other by 1.5-2.0 m high partitions when performing creative work that requires extreme mental stress or high concentration.
- The height of the bench work surface for adult use should be adjusted within 680-800 mm; if this is not possible, the height of the bench work surface should be 725 mm.
- Workplaces for PC users should be equipped with footrests with a width of at least 300 mm, a depth of at least 400 mm, a height adjustable to 150 mm, and an inclination angle of 20° on the support surface of the stand. The surface of the stand must be corrugated and have a 10mm high edge along the front edge.
- The keyboard should be placed on a desktop 100-300mm from the user-facing edge, or on a special height-adjustable work surface separate from the main desktop.
Rules for labor protection and safety measures are introduced in order to prevent accidents, ensure safe working conditions for workers and are mandatory for workers, managers, engineers and technicians.
According to the SanPiN 2.2.4.1329-03, personal protective equipment when working on a computer includes spectral computer glasses to improve image quality and Protection against excessive energy flows of visible light and for Prof. Glasses reduce eye fatigue by 25-30%.
They are recommended to be used by all operators when working more than 2 hours a day, and in case of visual impairment by 2 diopters or more - regardless of the duration of work.

10.3 Occupational safety

Working conditions - a set of factors of the working environment and the labor process that affect the performance and health of a person.

Safe working conditions - working conditions under which the impact on workers of harmful and (or) hazardous production factors is excluded or the levels of their impact do not exceed the established standards.

The purpose of labor protection - to minimize the likelihood of injury or illness of working personnel at maximum labor productivity.

Hazardous production factor (OPF) - is called such a production factor, the impact of which on the employee can lead to injury.

Harmful production factor (HFF) - such a production factor, the impact of which on the employee can lead to his illness or reduced disability.

GOST 12.0.003-2015 "*Hazardous and harmful production factors. Classification"* must be used to identify potential factors, that can effect on a worker (employee).

Table 20. - Potential hazardous and harmful production factors

Factors (**GOST 12.0.003-2015**)	Legislation documents
1. Lack or lack of natural light, insufficient illumination.	**SanPiN 2.2.1/2.1.1.1278-03** Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings;
2. Excessive levels of noise	**GOST 12.1.003-2014** Occupational safety standards system. Noise. General safety requirements
3. Electromagnetic fields.	**SanPiN 2.2.4.1329-03** Requirements for protection of personnel from the impact of impulse electromagnetic fields;
4. Physical overload (static-long-term preservation of a certain posture).	**GOST 12.2.032-78 SSBT.** Workplace when performing

	work while sitting General ergonomic requirements.

10.4 Lack or lack of natural light, insufficient illumination

Light sources can be both natural and artificial. The natural source of the light in the room is the sun, artificial light are lamps. In the working environment of the computer, there may be weather reasons or insufficient indoor lighting.

With long work in low illumination conditions and in violation of other parameters of the illumination, visual perception decreases, myopia, eye disease develops, and headaches appear.

According to the SanPiN 2.2.1/2.1.1.1278-03 standard, the illumination on the table surface in the area of the working document should be 300-500 lux. Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300 lux.

The brightness of the lamps of common light in the area with radiation angles from 50 to 90° should be no more than 200 cd/m, the protective angle of the lamps should be at least 40°. The ripple coefficient should not exceed 5%.

10.5 Excessive levels of noise

Noise can be generated by the fan of the computer processor, the fan of the computer graphics card, or it can be transmitted from the outside.

Noise worsens working conditions; have a harmful effect on the human body, namely, the organs of hearing and the whole body through the central nervous system. It results in weakened attention, deteriorated memory, decreased response, and increased number of errors in work.

When working on a PC, according to the GOST 12.1.003-2014 document the noise level in the workplace should not exceed 65 dB.

In order to study in a quiet environment, irrelevant applications of the computer should be closed to reduce computer power consumption, thereby reducing computer noise, and windows should also be closed to reduce environmental noise.

10.6 Increased electromagnetic field

In our case, according to the SanPiN 2.2.4.1329-03 standard, the sources of increased intensity of the electromagnetic field are a personal computer. 8 kA / m is considered acceptable. An hour's working day for an employee at his workplace, with the maximum permissible level of tension, should be no more than 8 kA / m, and the level of magnetic induction should be 10 mT. Compliance with these standards makes it possible to avoid the negative effects of electromagnetic radiation.

To reduce the level of the electromagnetic field from personal it is recommended to connect no more than two computers to one outlet, make a protective grounding, connect the computer to the outlet through an electric field neutralizer.

Sources of electromagnetic radiation in the workplace are system units and monitors of switched-on computers. To bring down exposure to such types of radiation, it is recommended to use such monitors, the radiation level is reduced, as well as to install protective screens and observe work and rest regimes.

According to the intensity of the electromagnetic field at a distance of 50 cm around the screen along the electrical component should be no more than:

- in the frequency range 5 Hz - 2 kHz - 25 V / m;
- in the frequency range 2 kHz - 400 kHz - 2.5 V / m.

The magnetic flux density should be no more than:

- in the frequency range 5 Hz - 2 kHz - 250 nT;
- in the frequency range 2 kHz - 400 kHz - 25 nT.

There are the following ways to protect against EMF:

- increase the distance from the source (the screen should be at least 50 cm from the user);
- the use of pre-screen filters, special screens and other personal protective equipment.

10.7 Physical overload

As a computer worker, you often need to sit in front of the computer for a long time to work. This static working posture will cause physical overload for a

long time. If you sit for a long time, your body will protest, and many people will have occupational diseases. If you don't pay attention to timely improvement, it will have a great impact on your health.

Wrist pain and decreased finger flexibility: Some friends who are engaged in writing work need to keep typing on the keyboard at the computer when they go to work. If the hand performs a single activity for a long time, the wrist joints, fingers and other parts may appear In the case of faint pain, the flexibility of the fingers will also decrease.

Lumbar and cervical vertebrae are often painful: when many people sit for a long time, they can maintain a straight posture at first, and the whole person can sit very upright, but with the development of time, many people will begin to hunched over. The neck will also stretch forward, and the lumbar spine, cervical spine and other parts of the human body will experience faint pain, and some people may even have problems such as lumbar disc herniation.

According to the GOST 12.2.032-78 SSBT standard, The design of the workplace and the relative position of all its elements (seats, controls, ways of displaying information, etc.) must conform to anthropometric, physiological and psychological requirements and the nature of the work.

It is recommended to prepare a comfortable small pillow at ordinary times and place the pillow on the chair, which will help relieve the pressure on the lumbar spine of the human body, and turn the neck more in leisure time, which can move the muscles and bones well.

Therefore, after hitting the keyboard for about an hour, you should stop and move your wrist joints. You can stretch your fingers together, which can also relax our arms.

For office workers, after sitting for a long time, the body is prone to some occupational diseases. Everyone should pay attention to avoid it in peacetime, and pay attention to replenishing the water in the body every day. The daily water intake is 2000 ml, which can be very good. Promote water circulation in the body, and get up and walk in moderation after sitting for a long time, which is more

conducive to the health of the body.

10.8 Environmental Safety

Presently section discusses the environmental impacts of the project development activities, as well as the product itself as a result of its implementation in production. The software product itself, developed during the implementation of the master's thesis, does not harm the environment either at the stages of its development or at the stages of operation. However, the funds required to develop and operate it can harm the environment.

There is no production in the laboratory. The waste produced in the premises, first of all, can be attributed to waste paper, plastic waste, defective parts of personal computers and other types of computers. Waste paper is recommended accumulate and transfer them to waste paper collection points for further processing. Place plastic bottles in specially designed containers.

Modern PCs are produced practically without the use of harmful substances hazardous to humans and the environment. Exceptions are batteries for computers and mobile devices. Batteries contain heavy metals, acids and alkalis that can harm the environment by entering the hydrosphere and lithosphere if not properly disposed of. For battery disposal it is necessary to contact special organizations specialized in the reception, disposal and recycling of batteries.

Fluorescent lamps used for artificial illumination of workplaces also require special disposal, because they contain from 10 to 70 mg of mercury [1], which is an extremely dangerous chemical substance and can cause poisoning of living beings, and pollution of the atmosphere, hydrosphere and lithosphere. The service life of such lamps is about 5 years, after which they must be handed over for recycling at special reception points. Legal entities are required to hand over lamps for recycling and maintain a passport for this type of waste. An additional method to reduce waste is to increase the share of electronic document management [1].

10.9 Emergency Safety

An emergency poses an immediate risk of significant harm to health, life, property or the environment. Preparing for emergencies is an important part of workplace health and safety program. An emergency situation is a situation in a certain territory that has developed as a result of an accident, hazardous natural phenomenon, catastrophe or other disaster, which may entail human casualties, damage to human health or the environment, significant material losses and violation of the living conditions of people. Emergency for the presented work space is a fire. This emergency can occur in the event of non-compliance with fire safety measures, violation of the technique of using electrical devices and PCs, violations of the wiring of electrical networks and a number of other reasons.
The working space provided for the performance of the WRC, according to NPB 105-03 [2], can be classified as category B (fire hazard).
The following reasons can be indicated as possible causes of a fire:

- short circuit.
- dangerous overload of networks, which leads to strong heating of live parts and ignition of insulation.
- start-up of equipment after incorrect and unqualified repairs.

To prevent emergencies, it is necessary to comply with fire safety rules in order to ensure the state of protection of employees and property from fire
To protect against short circuits and overloads, it is necessary to correctly select, install and use electrical networks and automation equipment.
To prevent the occurrence of fires, it is necessary to exclude the formation of a combustible environment, to monitor the use of non-combustible or hardly combustible materials in the construction and decoration of buildings.
It is necessary to carry out the following fire prevention measures:

- organizational measures related to the technical process, taking into account the fire safety of the facility (personnel briefing, training in safety rules, publication of instructions, posters, evacuation plans).

• operational measures that consider the operation of the equipment used (compliance with equipment operating standards, ensuring a free approach to equipment, maintaining conductor insulation in good condition).

• technical and constructive measures related to the correct placement and installation of electrical equipment and heating devices (compliance with fire safety measures when installing electrical wiring, equipment, heating, ventilation and lighting systems).

To increase the resistance of the working room to emergencies, it is necessary to install fire alarm systems that react to smoke and other combustion products, install fire extinguishers. Also, two times a year to conduct drills to practice actions in case of fire.

An evacuation plan is presented in the presented working room at the entrance, a fire alarm system is installed. The room is equipped with OU-2 type carbon dioxide fire extinguishers in the amount of 2 pieces per one working area. There is an electrical panel within the reach of workers, with the help of which it is possible to completely de-energize the working room. In the event of a fire, you must call the fire department by phone 101 and inform the place of the emergency, take measures to evacuate workers in accordance with the evacuation plan. In the absence of direct threats to health and life, make an attempt to extinguish the resulting fire with existing carbon dioxide fire extinguishers. In case of loss of control over the fire, it is necessary to evacuate after the employees according to the evacuation plan and wait for the arrival of the fire service specialists.

10.10 Conclusion for social responsibility

Every employee must carry out professional activities that take into account social, legal, environmental and cultural aspects, health and safety issues, take social responsibility for solutions, and be aware of the need for sustainable development.

This section includes the main issues and solutions for respecting employees' right to work, complying with labor safety, industrial safety, ecological and resource protection rules.

10.11 Reference for social responsibility

GOST R 52105-2003 Resources saving. Waste treatment.

NPB 105-03. Definition of categories of premises, buildings and outdoor.

References

1. Mohit Sharma.(2018).What Steps should one take while doing Data Preprocessing? https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa
2. A Guide to Logistic Regression in SAS(2019) https://communities.sas.com/t5/SAS-Communities-Library/A-Guide-to-Logistic-Regression-in-SAS/ta-p/564323#
3. Deepanshu Bhalla LOGISTIC REGRESSION ANALYSIS WITH SAS https://www.listendata.com/2013/04/logistic-regression-analysis-with-sas.html
4. Huang Shan, Gubin E. Data cleaning for data analysis. Томск, 2018г. - С. 387-389.
5. Onesmus Mbaabu. (09.04.2020) Introduction to Random Forest in Machine Learning [Electronic resource]. – URL: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning
6. The application of big data in the financial industry [Electronic resource]. – URL: http://c.biancheng.net/view/3736.html
7. Prashant Gupta. (18.05.2017) Decision Trees in Machine Learning [Electronic resource]. – URL: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

List of terms and abbreviations

SAS – Statistical analysis system

HIVE – The Apache Hive data warehouse software

LOG R – Logistic regression

AUC – Area Under the ROC Curve

ROC – Receiver Operating Characteristic Curve

N – Number of samples

GB – Good or bad

DT – Decision Trees

KNN – K-NearestNeighbor

SVM – Support Vector Machine

GNB – GaussianNB

RF – Random Forest

Appendix A Program cod for scorecard (GUI)

Code:

```
# -*- coding: utf-8 -*-

from PyQt5.QtWidgets import *

from PyQt5.QtWidgets import QApplication

from PyQt5.QtCore import QCoreApplication

import copy


app = QApplication([])

window = QMainWindow()


window.resize(1200, 1080)

window.move(500, 500)

window.setWindowTitle('CreditCard_GUI')


table = QTableWidget(window)

table.move(20, 20)

table.resize(800, 600)

table.setObjectName("sql_result_out")


x1_score = [2.0, -1.0, -1.0, -1.0, 0.0]

x2_score = [4.0, 2.0, 1.0, 0.0, -1.0, -1.0, -2.0, -5.0]

x4_score = [3.0, 1.0, 0.0, 0.0, -0.0, -1.0, -4.0]
```

```
  x19_score = [-0.0, -5.0, -6.0, 2.0]
  baseScore = 88.0

  ninf = float('-inf')
  pinf = float('inf')
  x1_cut = [ninf, 1, 2, 3, 5, pinf]
  x2_cut = [ninf, 23.0, 25.0, 28.0, 31.0, 35.0, 39.0,
46.0, pinf]
  x4_cut = [ninf, 12.0, 21.0, 33.0, 45.0, 72.0, 144.0,
pinf]
  x19_cut = [ninf, 1, 2, 5, pinf]

  score = [x1_score,x2_score,x4_score,x19_score]
  cut_t = [x1_cut,x2_cut,x4_cut,x19_cut]

  def compute_score(x):
      tot_score = baseScore
      cut_d = copy.deepcopy(cut_t)
      for j in range(len(cut_d)):
          cut_d[j].append(x[j])
          cut_d[j].sort()
          for i in range(len(cut_d[j])):
              if cut_d[j][i] == x[j]:
```

```
                tot_score = score[j][i-1] +tot_score
    return tot_score


PERS_H = QLineEdit(window)
PERS_H.resize(150, 30)
PERS_H.move(180, 700)
PERS_H_lable = QLabel(window)
PERS_H_lable.setText('PERS_H:')
PERS_H_lable.resize(150, 30)
PERS_H_lable.move(50, 700)


AGE = QLineEdit(window)
AGE.resize(150, 30)
AGE.move(530, 700)
AGE_lable = QLabel(window)
AGE_lable.setText('AGE:')
AGE_lable.resize(150, 30)
AGE_lable.move(400, 700)


TMJOB1 = QLineEdit(window)
TMJOB1.resize(150, 30)
TMJOB1.move(180, 780)
TMJOB1_lable = QLabel(window)
```

```
  TMJOB1_lable.setText('TMJOB1:')
  TMJOB1_lable.resize(150, 30)
  TMJOB1_lable.move(50, 780)

  cards_ = QLineEdit(window)
  cards_.resize(150, 30)
  cards_.move(530, 780)
  cards__lable = QLabel(window)
  cards__lable.setText('cards_:')
  cards__lable.resize(150, 30)
  cards__lable.move(400, 780)

  def calculate():
      #x = [PERS_H, AGE, TMJOB1, cards_]
      PERS_H_value = int(PERS_H.text())
      AGE_value = int(AGE.text())
      TMJOB1_value = int(TMJOB1.text())
      cards__value = int(cards_.text())
      list_x = [PERS_H_value, AGE_value, TMJOB1_value,
cards__value]
      int_x = [int(x) for x in list_x]
      scoreResults = compute_score(int_x)
```

```
        return scoreResults, PERS_H_value, AGE_value,
TMJOB1_value, cards__value

        #  print(PERS_H_value, AGE_value, TMJOB1_value,
cards__value)

        # print(int_x)

    def test():

        PERS_H_value = PERS_H.text()

        print(f'PERS_H_value:{PERS_H_value}')


    def showDialog():

        scoreValue, PERS_H_value, AGE_value, TMJOB1_value,
cards__value= calculate()

        dialog = QDialog()

        dialog.resize(800,500)


        Results = QLabel(dialog)

        Results.setScaledContents(True)

        Results.setText(f'Evaluate successfully! The score
of this customer is: {scoreValue}')

        #Results.move(50, 50)


        bto = QPushButton("OK",dialog)

        #bto.move(325,400)

        #bto.resize(150,60)
```

```
    bto.clicked.connect(dialog.close)

    vbox = QVBoxLayout()
    vbox.addWidget(Results)
    vbox.addWidget(bto)

    dialog.setLayout(vbox)

    dialog.setWindowTitle('Result')
    dialog.exec_()

bt1 = QPushButton('Evaluate', window)
bt1.move(930, 730)
bt1.resize(150, 55)
bt1.clicked.connect(showDialog)

btq = QPushButton('Quit', window)
btq.move(930, 930)
btq.resize(150, 55)
btq.clicked.connect(QCoreApplication.instance().quit)

window.show()
```

```
app.exec_()
```