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
Assessment of the credit risk of consumer loan (Оценка кредитного риска потребительского кредита)

УДК 004.65:004.4516336.774:005.334

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И	Ли Кэ		

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	Ли Кэ

Тема работы:

Assessment of the credit risk of consumer loan (Оценка кредитного риска потребительского кредита)	
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.).*	Разработка рейтинга кредитного риска с использованием программирования Python и создание программы, которая поможет кредиторам или банкам в принятии решений.

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. Аналитический обзор литературных Источников. 2. постановка задачи исследования. 3. разработка методики. 4. реализация методики. 5. выбор программного обеспечения. 6. обсуждение результатов выполненной работы. 7. финансовый менеджмент. 8. социальная ответственность. 9. заключение.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Скриншот программы. 2. UML диаграммы.

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	Ли Кэ		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»

Student:

Group	Full name
8PM0I	Li Ke

School	Information Technology and Robotics	**Division**	Information Technology
Degree	Master	**Educational Program**	09.04.04 Software Engineering

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 –107,410.4 rub - STR budget –172,171.7 rub.
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	Menshikova E.V.	PhD		

The task was accepted by the student:

Group	Full name	Signature	Date
8PM0I	Li Ke		

TASK FOR SECTION «SOCIAL RESPONSIBILITY»

Student:

Group	Name
8PM0I	Li Ke

School	Information Technology and Robotics	**Division**	Information Technology
Degree	Master	**Educational Program**	09.04.04 Software Engineering

Topic of FQW:

Assessment of the credit risk of consumer loan	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	- Data analysis on customer loan of bank - Using machine learning algorithm to create credit risk model - Working area: working on TPU dormitory, with desktop and personal laptop
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. - SP 2.4.3648-20. Sanitary and Epidemiological Requirements for Organizations of Education and Training, Recreation and Recreation of Children and Youth
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the sourse of factor, the impact on human's body 2.2 Suggest measures to reduce the impact of identified harmful and dangerous factors	Dangerous and harmful factors: - Insufficient illumination of workplace - Abnormal microclimatic parameters of the air - Increased level of noise - Electromagnetic fields - 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	-Impact of the object on the lithosphere: disposal of electrical components and waste (failed mouse, failed keyboard, luminescent lamps)
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	Li Ke		

Abstract

Final qualifying work 68 pages, 21 figures, 18 tables, 21 sources.

Keywords: data analysis, data preparation, data cleaning, logistic regression, random forest, credit risk model.

With the vigorous development of the world economy and the change from people's consumption concept, loans have become an important way for individuals to solve economic problems. This thesis is researched about how to reduce loan risk by using machine learning model. This thesis contains data preparation, data cleaning, model building, model accuracy test and a program. This program is created by python, and it need to input consumer’s personal information. And with this program bank and other institutions could predict the probability of a consumer whether repay the loan on time, and with the result bank could make a final decision. It helps bank to reduce the risk of loan.

Publications:

Li Ke. Assessment of credit risk by using big data tools / Li Ke, E. I. Gubin // Молодежь и современные информационные технологии сборник трудов XIX Международной научно-практической конференции студентов, аспирантов и молодых учёных, 21-25 марта 2022 г., г. Томск: / Национальный исследовательский Томский политехнический университет, Инженерная школа информационных технологий и робототехники ; ред. кол. А. Ю. Дёмин, Н. Г. Марков, В. Г. Спицын [и др.] . — Томск : Изд-во ТПУ , 2022 . — [С. 224-225]

CONTENT

Terms, abbreviations and conventions

SAS - Statistical Analysis System

SEMMA - Sample, Explore, Modify, Model, and Assess

EDA - Exploratory Data Analysis

LR - Logistic Regression

RF – Random Forests

GB - Good or Bad

ROC - Receiver Operating Characteristic Curve

AUC - Area Under the ROC Curve

GUI - Graphic User Interface

Introduction

With the vigorous development of the world economy and the change from people's consumption concept, loans have become an important way for individuals to solve economic problems. With the growing demand of people, the probability of loan defaults has also increased. In order to avoid loan defaults, banks and other financial institutions will evaluate or score the borrower’s credit risk when issuing loans, predict the probability of loan default, and make a judgment on whether to issue a loan based on the results. How to effectively evaluate and identify the potential default risk of borrowers before issuing loans is the basis and important link of credit risk management of financial institutions. Using a set of scientific models and systems to determine the risk of loan default can minimize the risk and reduce the risk of default.

This thesis mainly studies how to use the big data tools to analyze the historical loan data of financial institutions such as banks, and predict the possibility of loan default based on the machine learning model. This project mainly uses SAS and Python for analysis and modeling. Python and SAS have powerful data analysis and drawing capabilities. Python is also good at develop GUI.

In this work, I made a literature overview to describe the step to create the Credit Risk Model, including:

- Data description
- EDA
- Data preparation
- Feature selection
- Model building
- Model accuracy test
- Create graphic user interface (Credit Risk Calculator)

1. Project and research methods

1.1 Software development tools and programming technology

SAS (Statistical Analysis System), it is for all the need for data processing, analysis of the computer or non-computer user to provide an easy to learn, easy to use, complete and reliable software system. Its unique language is characterized by simplicity, ease of learning and strong pertinence in sentences.

SAS has been widely used in finance, insurance, medicine, public health, STD prevention, telecommunications, transportation, customs, government, universities and research institutes, market research, agriculture, manufacturing and other fields. It is also the only commercial software developed by the pharmaceutical industry to provide statistical analysis for the development and evaluation of drugs. According to statistics, more than 75% of large enterprises and companies are using SAS as a BI analysis tool to make in-depth analysis and prediction for the development of the company [1].

Python language is a very simple and easy to learn, and it suitable for beginners as an entry language. Python has very mature libraries and active communities in data analysis and interaction, exploratory computing, and data visualization, making python an important solution for data processing tasks. In terms of data processing and analysis, python has a series of excellent libraries and tools such as Numpy, Pandas, Matplotlib, Scikit-learn, etc. In particular, Pandas can be said to have unparalleled advantages in processing medium-sized data.

Python is not only powerful in data analysis, but also has a wide range of applications in many fields such as crawler, web, automated operation and maintenance, and even games. This makes it possible for companies to use one technology to complete all services, which facilitates business integration between various technology groups [2].

In this paper, we used both SAS and Python to prepare and analyze data.

1.2 Data mining methodology

SEMMA is an acronym that stands for Sample, Explore, Modify, Model, and Assess. It is a list of sequential steps developed by SAS Institute [3].

S: Sample (collect data)

Based on needs, collect data that can solve problems in a targeted manner. Commonly used collection methods are:

- Questionnaire
- Database queries
- Laboratory tests
- Records of equipment

E: Explore (data exploration)

Through data exploration, in order to further understand the data. Commonly used exploration directions are:

- Distribution proportions of discrete variables
- Distribution patterns of continuous variables
- Anomalies and missing data
- Feature selection

M: Modify (data correction)

Data correction for further analysis and modeling. Commonly used correction methods are:

- Data type conversion
- Consistent handling of data
- Handling of outliers and missing values
- Transformation of data form

M: Model (data modeling)

Data modeling, which focuses on the prediction of unknown events. Commonly used models are:

- Supervised predictive models (regression, decision trees, KNN, etc.)
- Supervised discriminative models (Logistic, Bayesian, ensemble, etc.)
- Unsupervised models (Kmeans clustering, hierarchical clustering, density clustering, etc.)
- Semi-supervised models (association rules, etc.)

A: Assess (modular assessment)

Model evaluation, test the stability and practicality of the model. Commonly used inspection methods are:

- Confusion matrix
- ROC curve [4]

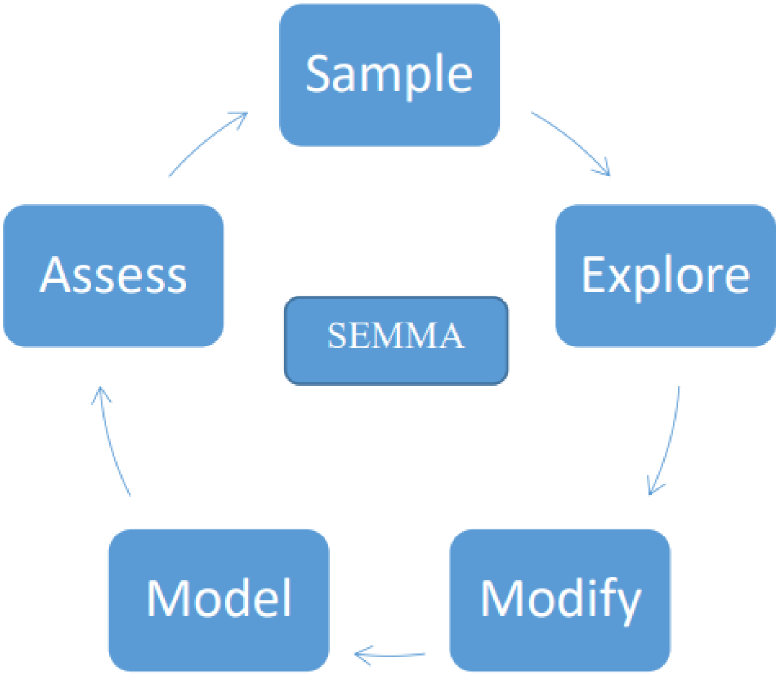


Figure 1.1 Method SEMMA

1.3 Data description

This dataset is historical data on creditworthiness. It contains 24 variables, and 3000 observations. GB is target variable, it means good or bad borrowers, 0 means good and 1 means bad. Bad borrowers are those borrowers who have not made the planned loan payments within 90 days [5]. Other variables contain age, amounts of children, income, region, etc. They are related to personal information of consumers.

Table 1 shows data description of this dataset.

Table 1. Variables of dataset

Variable name	Defination	Variable type
TITLE	The nature of homeownership	Categorical
CHILDREN	Amounts of children	Numeric
PERS_H	Number of people in the household	Numeric
AGE	Age	Numeric
TMADD	Number of months of residence at the current place of residence	Numeric
TMJOB1	Number of months in the current job	Numeric
TEL	Number of contact phone numbers	Numeric
NMBLOAN	The number of loans in this bank	Numeric
FINLOAN	No unpaid loans	Binary
INCOME	Income (per week in euros)	Numeric
EC_CARD	Possession of a bank card	Binary
INC	Salary	Numeric
INC1	Division into 5 categories according to the level of wages	Categorical
BUREAU	Credit risk class as assessed by the credit bureau	Categorical
LOANS	Number of loans outside the bank	Numeric
REGN	Region of residence	Categorical
CASH	Loan requested	Numeric
PRODUCT	Purpose of the loan	Categorical
RESID	Tenant or home owner	Categorical
NAT	Nationality	Categorical
PROF	Industry	Categorical
CAR	Vehicle type	Categorical
CARDS	Credit card type	Categorical

GB	Target variable /Good-Bad	Binary

1.4 Exploratory data analysis

EDA (Exploratory Data Analysis), is to understand the data set, understand the relationship between variables and the relationship between variables and predicted values, so as to help us better perform feature engineering and build models later. It is a very important step in data mining.

We could see age of consumers is mainly distributed from 20 to 40 years old in histogram of age.

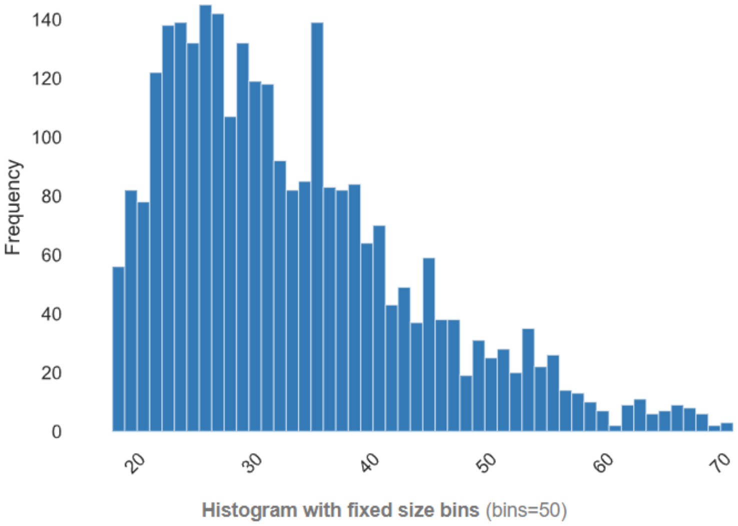


Figure 1.2 Histogram of age

The amount that the consumer wants to borrow is mainly distributed from 0-5000, also we could see in this feature has outliers on histogram.

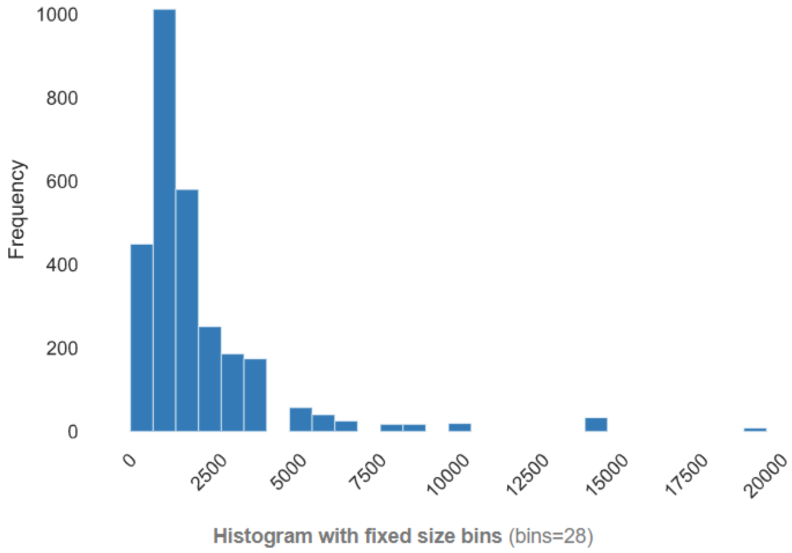


Figure 1.3 Histogram of cash

The income of consumer is mainly distributed from 0-4000, also it has outliers.

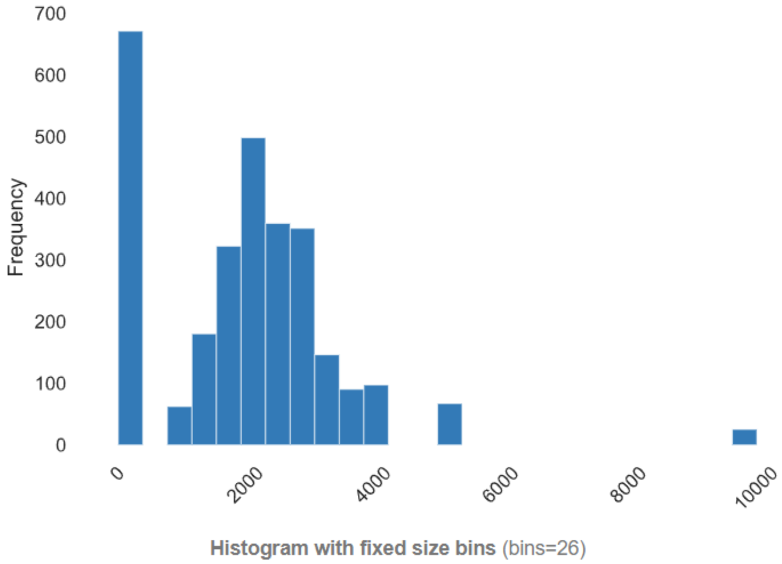


Figure 1.4 Histogram of income

1.5 Data preparation

When we want to analyze data, we first need to prepare the data. Data preparation is the process of cleaning and transforming raw data before processing and analysis. This is an important step before processing, and usually involves reformatting the data, correcting it, and merging datasets to enrich the data.

The first step is data cleaning. Data cleaning is a very important step of data preparation, if we ignore this step, we will get wrong result in final. In general, there are 6 problems we need to care about, they are missing data, mistake of data, outliers of data, duplicate cases, multicollinearity of data and digitalization of data [6].

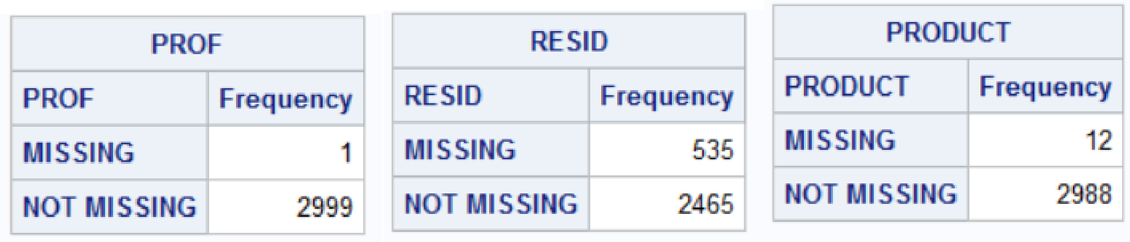
In this part, we used SAS to clean data.

At first, we need to import our dataset:

```
proc import out=sasuser.accepts
datafile="C:\Users\LIKE\Desktop\like sas\accepts.xls"
dbms=xls REPLACE;
run;
```

This code is using to find missing values:

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



PROF	
PROF	Frequency
MISSING	1
NOT MISSING	2999

RESID	
RESID	Frequency
MISSING	535
NOT MISSING	2465

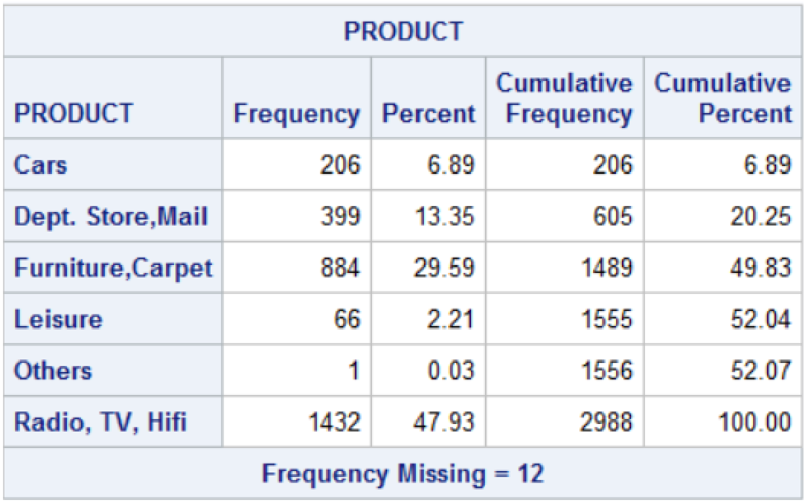
PRODUCT	
PRODUCT	Frequency
MISSING	12
NOT MISSING	2988

Figure 1.5 Missing values of data set

In this Figure 2, we can see PROF has 1 missing value, RESID has 535 missing values, and PRODUCT has 2 missing values.

Their frequency:

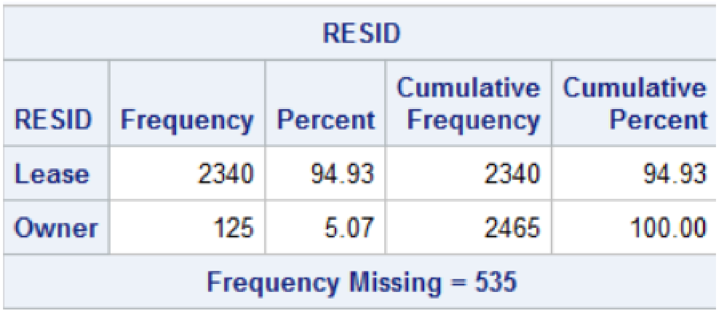
```
proc freq data=sasuser.accepts;
table product resid nat prof car cards gb;
title 'Descripitive statistics of char';
run;
```



PRODUCT

PRODUCT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cars	206	6.89	206	6.89
Dept. Store,Mail	399	13.35	605	20.25
Furniture,Carpet	884	29.59	1489	49.83
Leisure	66	2.21	1555	52.04
Others	1	0.03	1556	52.07
Radio, TV, Hifi	1432	47.93	2988	100.00

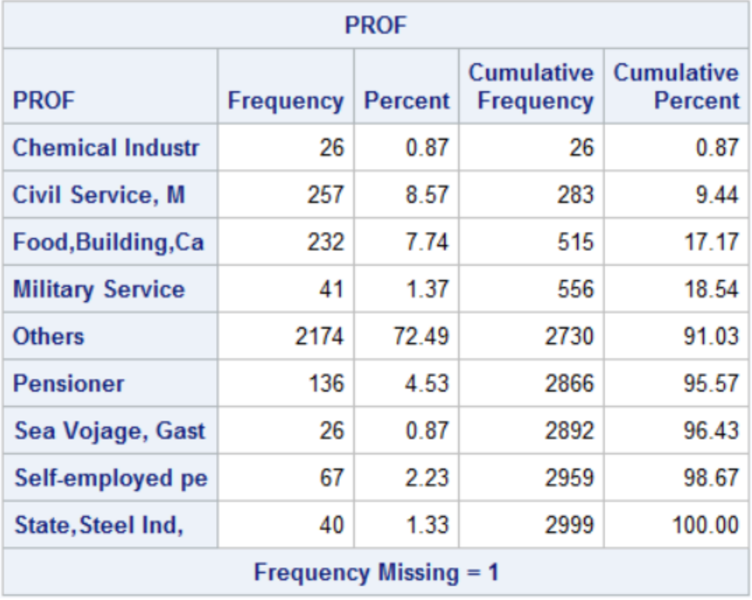
Frequency Missing = 12



RESID

RESID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Lease	2340	94.93	2340	94.93
Owner	125	5.07	2465	100.00

Frequency Missing = 535



PROF				
PROF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Chemical Industr	26	0.87	26	0.87
Civil Service, M	257	8.57	283	9.44
Food,Building,Ca	232	7.74	515	17.17
Military Service	41	1.37	556	18.54
Others	2174	72.49	2730	91.03
Pensioner	136	4.53	2866	95.57
Sea Vojage, Gast	26	0.87	2892	96.43
Self-employed pe	67	2.23	2959	98.67
State,Steel Ind,	40	1.33	2999	100.00
Frequency Missing = 1				

Figure 1.6 Frequency of product, resid and prof

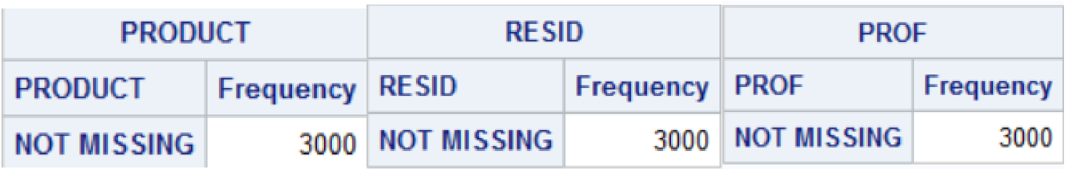
Change missing values to most common value:

```
data sasuser.accepts1;   /*new dataset*/
set sasuser.accepts;
if product='' then product='Radio, TV, Hifi';
if resid='' then resid='Lease';
if prof='' then prof='Others';
run;
```

Use same code to check it:

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

```
RUN;
```



PRODUCT		RESID		PROF	
PRODUCT	Frequency	RESID	Frequency	PROF	Frequency
NOT MISSING	3000	NOT MISSING	3000	NOT MISSING	3000

Figure 1.7 Check result

The next step is finding outliers:

```
proc univariate data=sasuser.accepts1 robustscale plot;
var inc;
run;
```

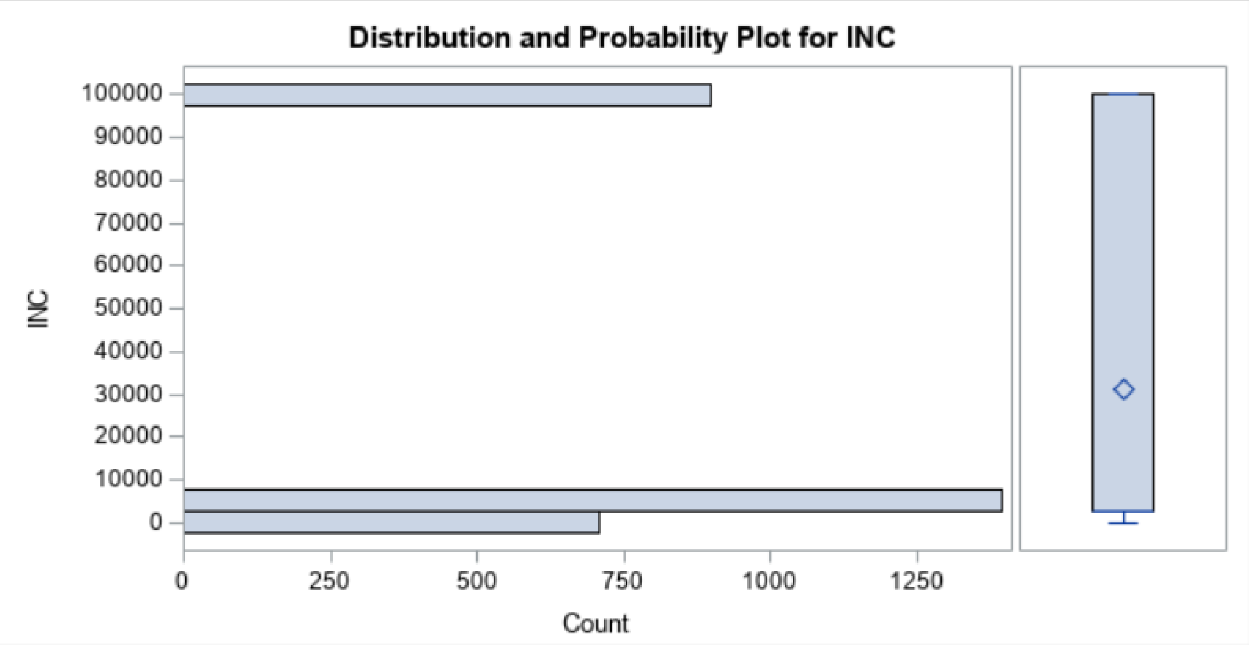


Figure 1.8 Find outliers

We can see that there are outliers here in column INC. We have omitted other codes for other columns here.

Delete all outliers:

```
data sasuser.accepts2;
set sasuser.accepts1;
if children=23 then delete;
if tmadd=999 then delete;
if TMJOB1=999 then delete;
if income=100000 then delete;
if cash=100000 then delete;
run;
```

Check it again:

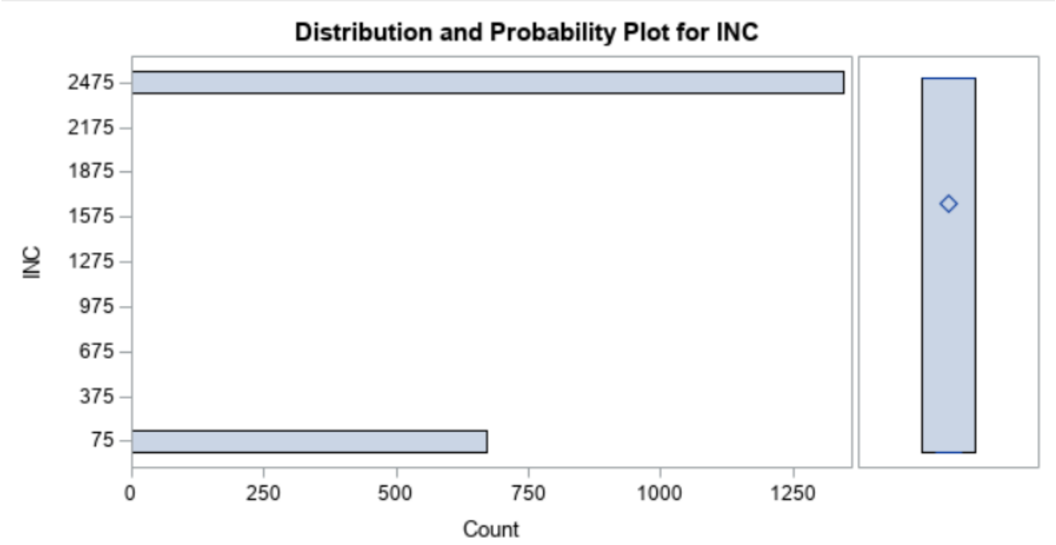
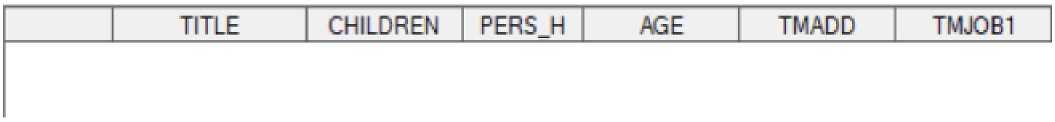


Figure 1.9 Check outliers of inc

The third step is finding duplicate values:

```
proc sort data=sasuser.accepts2 nounikey out=dup;/*find
duplicates*/
by title children pers_h age tmadd tmjob1 tel nmbloan
finloan income ec_card inc1 bureau loans regn cash product
resid nat prof car cards;
run;
```



	TITLE	CHILDREN	PERS_H	AGE	TMADD	TMJOB1

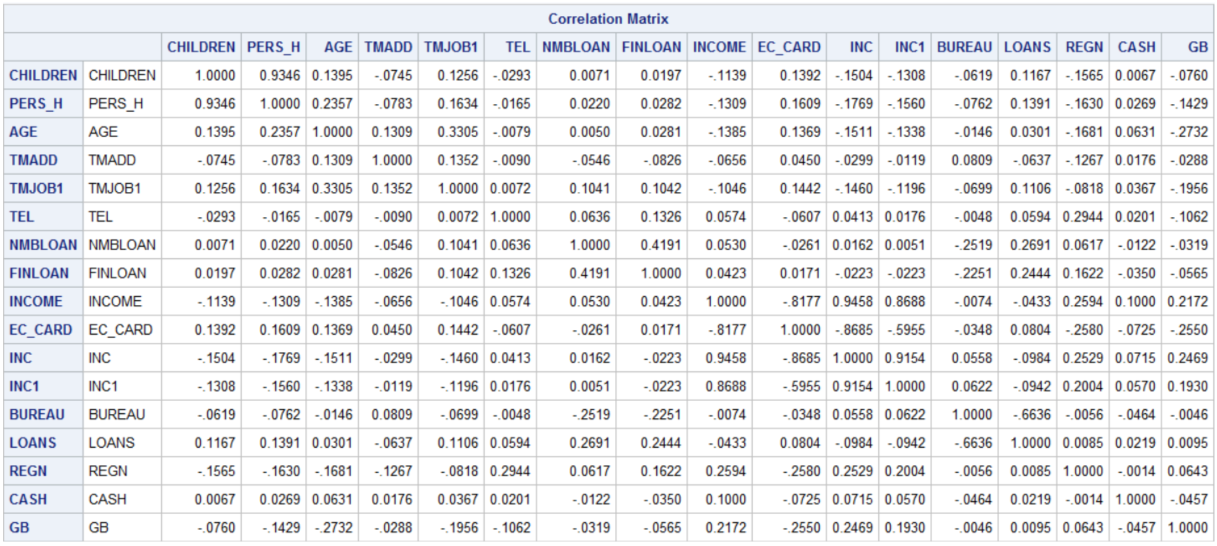
Figure 1.10 Part of table dup

This code could find duplicate values, and put them in a new table, now this table is empty, so no duplicate value in our dataset. If we have duplicate values, we can use this code:

```
    proc sort data=sasuser.accepts2 nodupkey out=NotDuplicate;
    by title children pers_h age tmadd tmjob1 tel nmbloan
finloan income ec_card inc1 bureau loans regn cash product
resid nat prof car cards;
    run;
```

The fourth step is checking multicollinearity:

```
proc princomp data=sasuser.accepts2 outstat=c_stat ;
run;
```



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	0.9346	0.1395	-.0745	0.1256	-.0293	0.0071	0.0197	-.1139	0.1392	-.1504	-.1308	-.0619	0.1167	-.1565	0.0067	-.0760
PERS_H	PERS_H	0.9346	1.0000	0.2357	-.0783	0.1634	-.0165	0.0220	0.0282	-.1309	0.1609	-.1769	-.1560	-.0762	0.1391	-.1630	0.0269	-.1429
AGE	AGE	0.1395	0.2357	1.0000	0.1309	0.3305	-.0079	0.0050	0.0281	-.1385	0.1369	-.1511	-.1338	-.0146	0.0301	-.1681	0.0631	-.2732
TMADD	TMADD	-.0745	-.0783	0.1309	1.0000	0.1352	-.0090	-.0546	-.0826	-.0656	0.0450	-.0299	-.0119	0.0809	-.0637	-.1267	0.0176	-.0288
TMJOB1	TMJOB1	0.1256	0.1634	0.3305	0.1352	1.0000	0.0072	0.1041	0.1042	-.1046	0.1442	-.1460	-.1196	-.0699	0.1106	-.0818	0.0367	-.1956
TEL	TEL	-.0293	-.0165	-.0079	-.0090	0.0072	1.0000	0.0636	0.1326	0.0574	-.0607	0.0413	0.0176	-.0048	0.0594	0.2944	0.0201	-.1062
NMBLOAN	NMBLOAN	0.0071	0.0220	0.0050	-.0546	0.1041	0.0636	1.0000	0.4191	0.0530	-.0261	0.0162	0.0051	-.2519	0.2691	0.0617	-.0122	-.0319
FINLOAN	FINLOAN	0.0197	0.0282	0.0281	-.0826	0.1042	0.1326	0.4191	1.0000	0.0423	0.0171	-.0223	-.0223	-.2251	0.2444	0.1622	-.0350	-.0565
INCOME	INCOME	-.1139	-.1309	-.1385	-.0656	-.1046	0.0574	0.0530	0.0423	1.0000	-.8177	0.9458	0.8688	-.0074	-.0433	0.2594	0.1000	0.2172
EC_CARD	EC_CARD	0.1392	0.1609	0.1369	0.0450	0.1442	-.0607	-.0261	0.0171	-.8177	1.0000	-.8685	-.5955	-.0348	0.0804	-.2580	-.0725	-.2550
INC	INC	-.1504	-.1769	-.1511	-.0299	-.1460	0.0413	0.0162	-.0223	0.9458	-.8685	1.0000	0.9154	0.0558	-.0984	0.2529	0.0715	0.2469
INC1	INC1	-.1308	-.1560	-.1338	-.0119	-.1196	0.0176	0.0051	-.0223	0.8688	-.5955	0.9154	1.0000	0.0622	-.0942	0.2004	0.0570	0.1930
BUREAU	BUREAU	-.0619	-.0762	-.0146	0.0809	-.0699	-.0048	-.2519	-.2251	-.0074	-.0348	0.0558	0.0622	1.0000	-.6636	-.0056	-.0464	-.0046
LOANS	LOANS	0.1167	0.1391	0.0301	-.0637	0.1106	0.0594	0.2691	0.2444	-.0433	0.0804	-.0984	-.0942	-.6636	1.0000	0.0085	0.0219	0.0095
REGN	REGN	-.1565	-.1630	-.1681	-.1267	-.0818	0.2944	0.0617	0.1622	0.2594	-.2580	0.2529	0.2004	-.0056	0.0085	1.0000	-.0014	0.0643
CASH	CASH	0.0067	0.0269	0.0631	0.0176	0.0367	0.0201	-.0122	-.0350	0.1000	-.0725	0.0715	0.0570	-.0464	0.0219	-.0014	1.0000	-.0457
GB	GB	-.0760	-.1429	-.2732	-.0288	-.1956	-.1062	-.0319	-.0565	0.2172	-.2550	0.2469	0.1930	-.0046	0.0095	0.0643	-.0457	1.0000

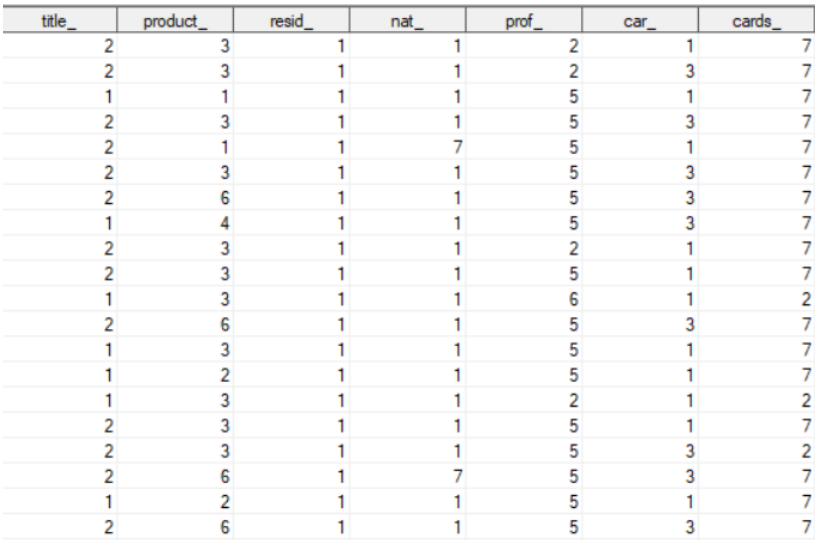
Figure 1.11 Correlation Matrix

The fifth step is Digitalization of data:

```
data sasuser.accepts3;
set sasuser.accepts2;
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='Furniture,Carpet' then product_=3;
if product='Leisure'                 then product_=4;
if product='Others'                 then product_=5;
if product='Radio, TV, Hifi'     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='Others'               then nat_=5;
if nat='Spanish/Portugue'     then nat_=6;
if nat='Turkish'              then nat_=7;
if nat='Yugoslav'             then nat_=8;
if prof='Chemical Industr'     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='State,Steel Ind,'     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	3	1	1	2	1	7
2	3	1	1	2	3	7
1	1	1	1	5	1	7
2	3	1	1	5	3	7
2	1	1	7	5	1	7
2	3	1	1	5	3	7
2	6	1	1	5	3	7
1	4	1	1	5	3	7
2	3	1	1	2	1	7
2	3	1	1	5	1	7
1	3	1	1	6	1	2
2	6	1	1	5	3	7
1	3	1	1	5	1	7
1	2	1	1	5	1	7
1	3	1	1	2	1	2
2	3	1	1	5	1	7
2	3	1	1	5	3	2
2	6	1	7	5	3	7
1	2	1	1	5	1	7
2	6	1	1	5	3	7

Figure 1.12 Digitalization of data

Digitization is the process of converting information into a digital format.

Finally, we need to set two datasets, one is train data, the other is test data. We use train data to analyze, and we use test data to check our result.

Data partition:

```
data sasuser.accepts_part;
set sasuser.accepts3;
label x='Random number';
x=ranuni(int(time()));
run;
data sasuser.accepts_test;
set sasuser.accepts_part;
if x>0.75;
```

```
run;
data sasuser.accepts_train;
set sasuser.accepts_part;
if x<=0.75;
run;
```

1.6 Random forest algorithm

Random forest is a classifier that contains multiple decision trees. The random forest algorithm was developed by Leo Breiman and Adele Cutler. Random forest, as the name implies, is to build a forest in a random way. The forest consists of many decision trees, and there is no relationship between these decision trees [7].

When we meet the classification task, with new input samples enter, each decision tree in the forest will be judged and classified separately. Each decision tree will get its own classification result. Which of the classification results of the decision tree has the most classification, then the random forest will regard this result as the final result [8].

Random forests can be used in many places:

1.Classification of Discrete Values

2.Regression on Continuous Values

3.Unsupervised Learning Clustering

4.Outlier detection

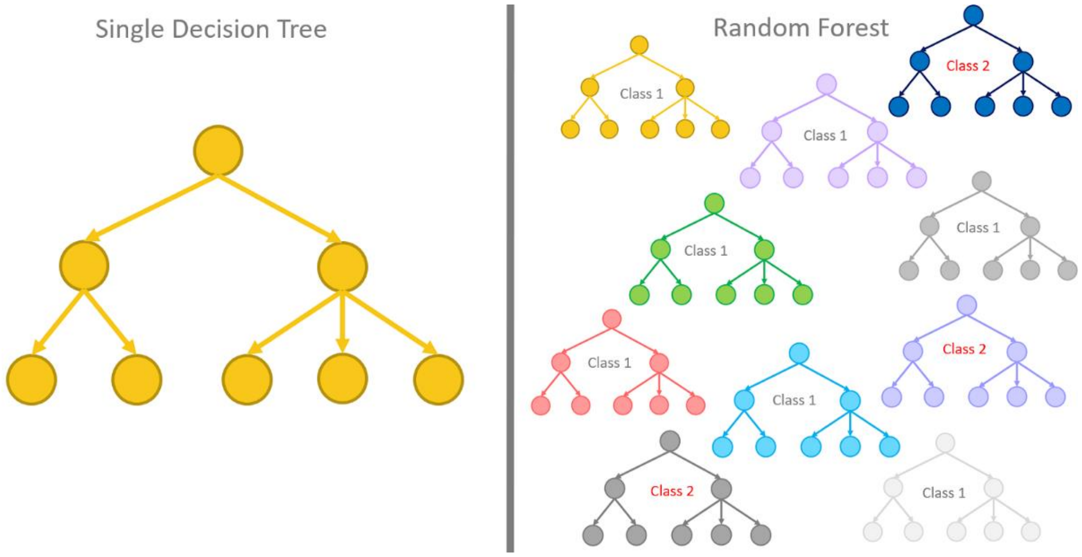


Figure 1.13 Decision tree and random forest

Advantage

It can produce very high-dimensional (many features) data without dimensionality reduction and feature selection.

It can judge the importance of features.

Interaction between different features can be judged.

Less prone to overfitting.

The training speed is relatively fast, and it is easy to make a parallel method.

Simpler to implement.

For imbalanced datasets, it can balance the error.

Accuracy can still be maintained if a significant portion of the features are missing.

Disadvantage

Random forests have been shown to overfit on some noisy classification or regression problems.

For data with attributes with different values, attributes with more value divisions will have a greater impact on random forests, so the attribute weights produced by random forest on such data are not credible.

1.7 Model building

We use random forest algorithm to create a model first by using Scikit-learn library from Python. Scikit-learn is the most popular library in machine learning area. This project uses *sklearn.ensemble.RandomForestClassifier* in Python to build a random forest model.

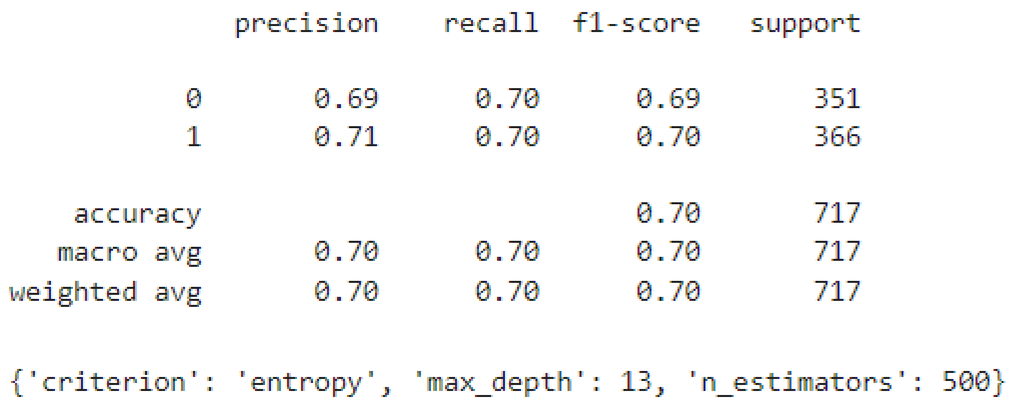
When we create model, we need to adjust parameter to get the better result. Here we used *GridSearchCV*.

GridSearchCV can actually be split into two parts, *GridSearch* and CV, namely grid search and cross-validation. Both names are very easy to understand. Grid search, searching for parameters, that is, within the specified parameter range, adjust the parameters in turn by step size, use the adjusted parameters to train the learner, and find the parameter with the highest accuracy on the validation set from all parameters. This is actually a training and comparison process.

GridSearchCV can guarantee to find the parameters with the highest accuracy within the specified parameter range, but this is also the defect of grid search. It requires to traverse all possible combinations of parameters, which is very time-consuming in the face of large data sets and multiple parameters.

Now we create model:

```
param_grid={'n_estimators':[500],
            'max_depth':[13],
            'criterion':['entropy']
}
rfc=RandomForestClassifier(random_state=0)
rfc_cv=GridSearchCV(estimator=rfc,param_grid=param_grid,cv=5)
rfc_cv.fit(X_train,y_train)
predict_test=rfc_cv.predict(X_test)
print(metrics.classification_report(predict_test,y_test))
```



```
              precision    recall  f1-score   support

           0       0.69      0.70      0.69       351
           1       0.71      0.70      0.70       366

    accuracy                           0.70       717
   macro avg       0.70      0.70      0.70       717
weighted avg       0.70      0.70      0.70       717

{'criterion': 'entropy', 'max_depth': 13, 'n_estimators': 500}
```

Figure 1.14 Result of model

In Figure we could see we have 70% accuracy, also we could see best parameters.

1.8 Model accuracy test

ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) value are often used to evaluate the quality of a binary classifier.

ROC curve

When it comes to the ROC curve, we must first explain two concepts: FPR (false positive rate), TPR (true positive rate). They are both an evaluation metric for classification tasks.

For a binary classification task (assuming that 1 represents a positive class and 0 represents a negative class), for a sample, there are four types of classification results:

The category is actually 1 and is divided into 0, FN (False Negative).

The category is actually 1 and is divided into 1, TP (True Positive).

The category is actually 0 and is divided into 1, FP (False Positive).

The category is actually 0 and is divided into 0, TN (True Negative).

And FPR (False Positive Rate) = FP / (FP + TN), that is, the proportion of negative class data is divided into positive class.

TPR (True Positive Rate) = TP / (TP + FN), that is, the proportion of positive class data is divided into positive class.

For the sample data, we use the classifier to classify it. The classifier will give the probability that each data is a positive example. We can set a threshold for this. When a sample is judged as a positive example, the probability is greater than this When the threshold is used, the sample is considered to be a positive example, and if it is less than a negative example, then we can get a (TPR, FPR) pair by calculation, that is, a point on the image. By continuously adjusting this threshold, we can get several points to draw a curve.

AUC

AUC, (Area Under Curve), is defined as the area under the ROC curve. Obviously, this area is less than 1, and because the ROC curve is generally above the line y=x, the AUC is generally between 0.5 and 1. The AUC value is used as the evaluation criterion because in many cases the ROC curve does not clearly indicate which classifier is better, and as a value, the classifier with a larger AUC is better.

The meaning of AUC is the probability that when a positive sample and a negative sample are randomly selected, the positive sample is ranked ahead of the negative sample according to the score calculated by the current classifier.

Criteria for judging the quality of a classifier (prediction model) from AUC:

AUC = 1, it is a perfect classifier, when using this prediction model, there is at least one threshold to get a perfect prediction. In the vast majority of prediction cases, there is no perfect classifier.

0.5 < AUC < 1, better than random guessing. This classifier (model) can have predictive value if the threshold is properly set.

AUC = 0.5, the following machine guesses the same (for example: losing a copper plate), the model has no predictive value.

AUC < 0.5, worse than random guessing; but better than random guessing as long as it always works against prediction.

Since there are so many evaluation criteria, why use ROC and AUC? Because the ROC curve has a good feature: when the distribution of positive and negative samples in the test set changes, the ROC curve can remain unchanged. In actual datasets, class imbalance often occurs, that is, there are many more negative samples than positive samples (or vice versa), and the distribution of positive and negative samples in the test data may also change over time [9].

In figure we could see AUC=0.748, and it means this model is not bad.

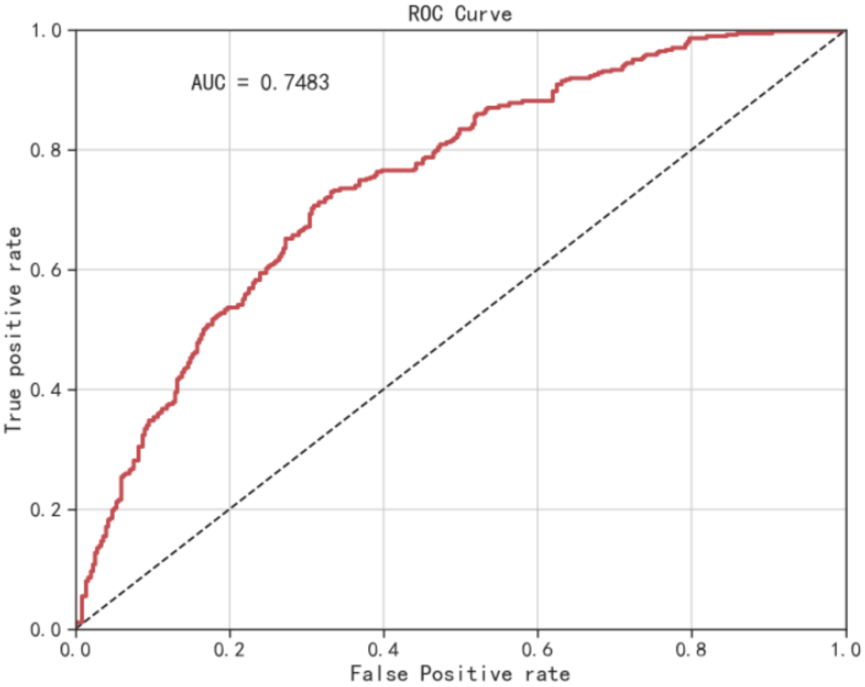


Figure 1.15 ROC curve for test data of random forest (by python)

Meanwhile I created a logistic regression model on SAS, we could see AUC =0.7225, so random forest model is a little better than logistic regression model on this dataset.

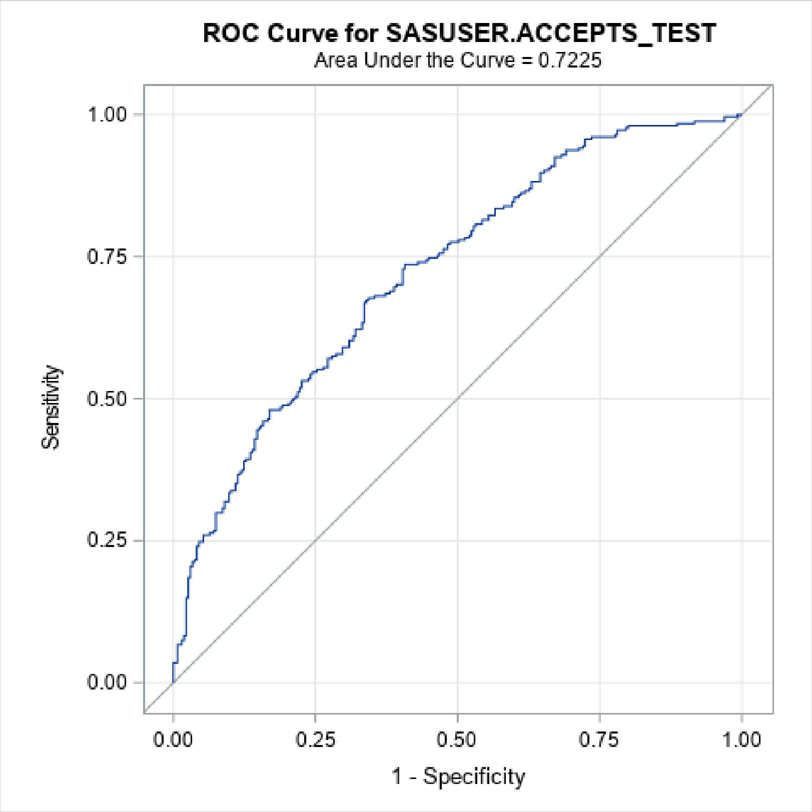


Figure 1.16 ROC curve for test data of logistic regression (by SAS)

1.9 Feature selection

The feature selection method based on feature importance is one of the most popular methods in machine learning. The principle is to select the importance of features through integrated models such as random forest.

Summarize the significance of feature selection:

1.Remove features that are not very relevant to the target variable

2.Eliminates variables due to linear correlation and avoids feature redundancy

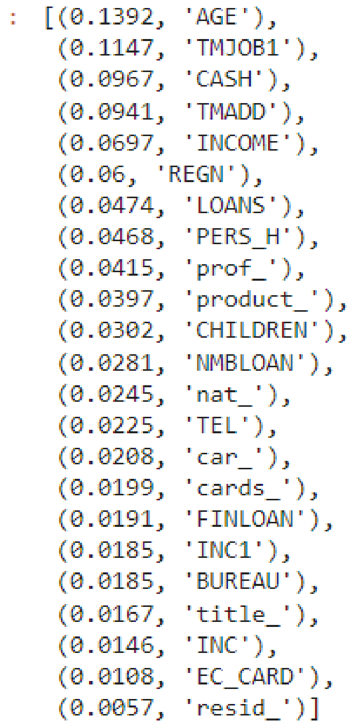
3.Reduce the burden of post-validation, deployment, and monitoring

4.Guaranteed interpretability of variables

In this dataset, we have 23 features about personal information of consumers, but some features are not important to the final result, so it is necessary to do a feature selection to select important features, so we could know which feature we need to consider at first when we want to loan to consumers.

After we created model, we coule use *feature_importances_ to* select important features.

```
imp=rfc_cv.best_estimator_.feature_importances_
imp
names=[]
for i in df1.columns:
    names.append(i)
feature_weights=sorted(zip(map(lambda x: round(x, 4), imp), names), reverse=True)
feature_weights
```



```
: [(0.1392, 'AGE'),
   (0.1147, 'TMJOB1'),
   (0.0967, 'CASH'),
   (0.0941, 'TMADD'),
   (0.0697, 'INCOME'),
   (0.06, 'REGN'),
   (0.0474, 'LOANS'),
   (0.0468, 'PERS_H'),
   (0.0415, 'prof_'),
   (0.0397, 'product_'),
   (0.0302, 'CHILDREN'),
   (0.0281, 'NMBLOAN'),
   (0.0245, 'nat_'),
   (0.0225, 'TEL'),
   (0.0208, 'car_'),
   (0.0199, 'cards_'),
   (0.0191, 'FINLOAN'),
   (0.0185, 'INC1'),
   (0.0185, 'BUREAU'),
   (0.0167, 'title_'),
   (0.0146, 'INC'),
   (0.0108, 'EC_CARD'),
   (0.0057, 'resid_')]
```

Figure 1.17 Feature importance

We can see feature importance from high at this list. We could know key factors of a customer are age, working time, cash, income, region, etc. Top 3 features are age, number of months in the current job, loan requested. That means these features we will consider first when we want to loan to customers.

1.10 Credit risk calculator (GUI)

After Feature selection, we selected top 8 features to developed a software to predict a consumer. Pyqt5 is a good GUI library in python, and we used this to design a credit risk calculator.

In this program, we could input information of consumer, and we click button ‘Predict’, then we will get a prediction, also we could see the probability to be a good or bad consumer. Also, we could click ‘Log’ to check historical predict result.

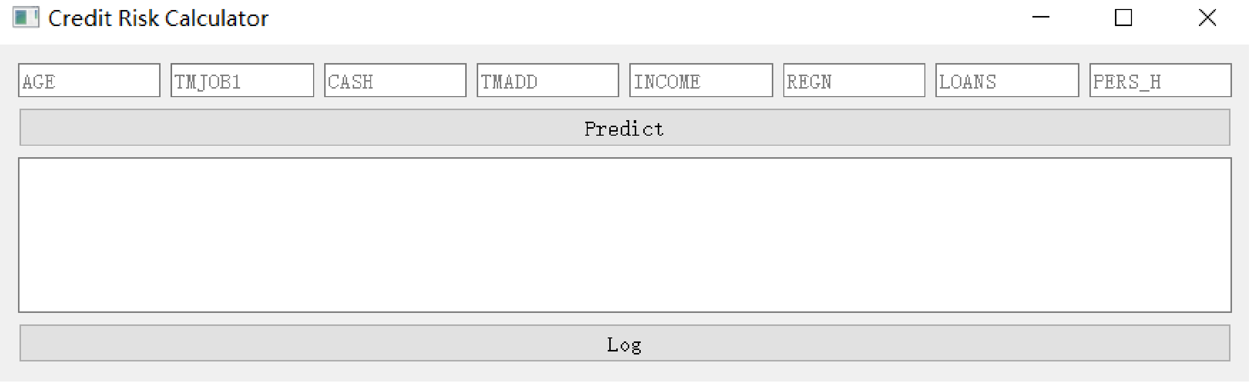
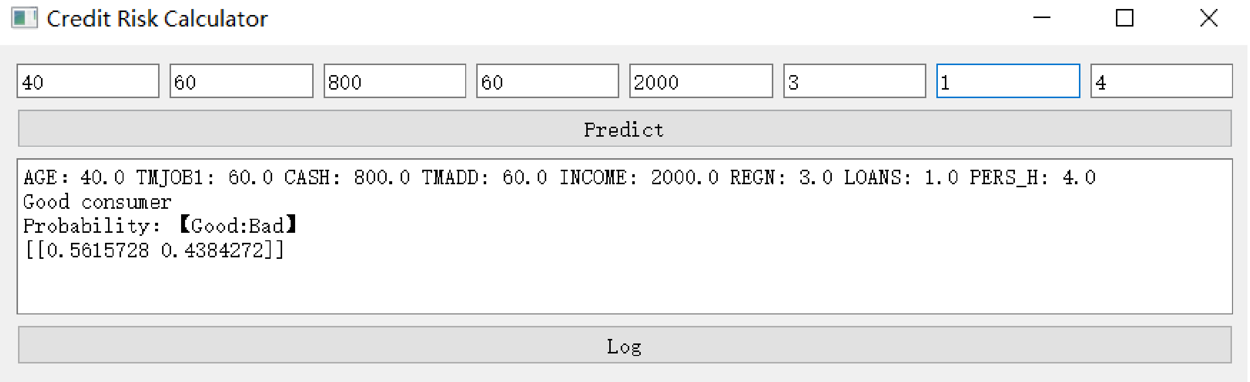


Figure 1.18 Credit risk calculator

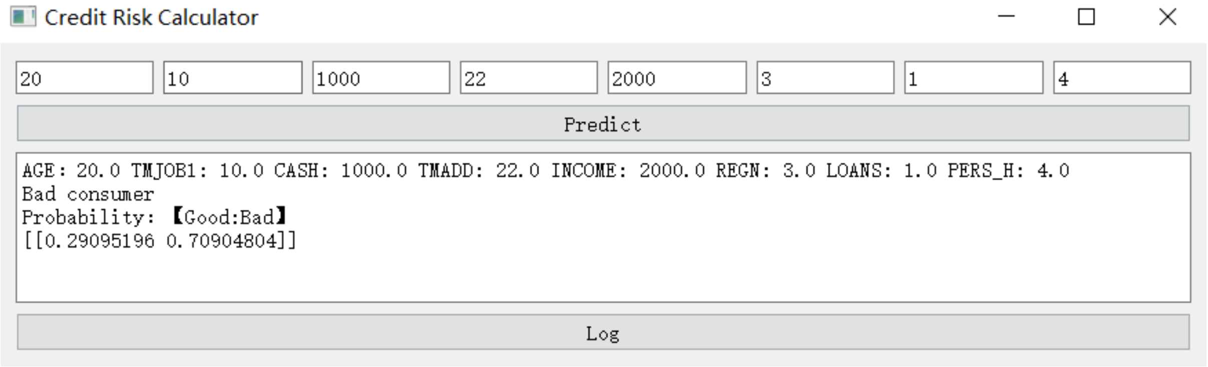
Here are 2 examples: prediction is good consumer and prediction is bad consumer. We could make a decision according to the result.



40 | 60 | 800 | 60 | 2000 | 3 | 1 | 4

```
AGE: 40.0 TMJOB1: 60.0 CASH: 800.0 TMADD: 60.0 INCOME: 2000.0 REGN: 3.0 LOANS: 1.0 PERS_H: 4.0
Good consumer
Probability: 【Good:Bad】
[[0.5615728 0.4384272]]
```

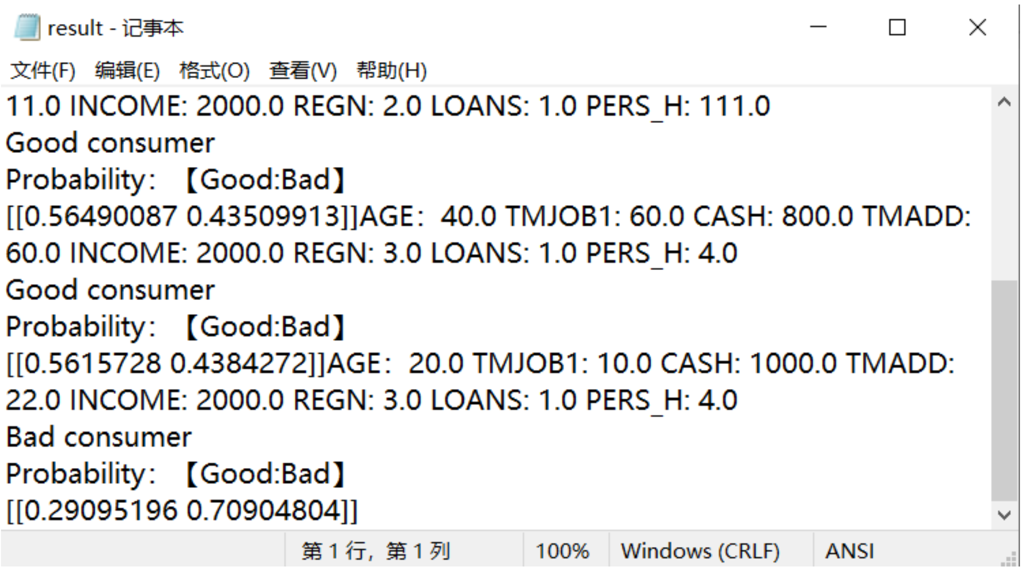
Figure 1.19 Credit risk calculator (good consumer)



```
AGE: 20.0 TMJOB1: 10.0 CASH: 1000.0 TMADD: 22.0 INCOME: 2000.0 REGN: 3.0 LOANS: 1.0 PERS_H: 4.0
Bad consumer
Probability: 【Good:Bad】
[[0.29095196 0.70904804]]
```

Figure 1.20 Credit risk calculator (bad consumer)

Here we could see historical predict result, so when we close the program unexpectedly, we could see our result in a txt file.



```
11.0 INCOME: 2000.0 REGN: 2.0 LOANS: 1.0 PERS_H: 111.0
Good consumer
Probability: 【Good:Bad】
[[0.56490087 0.43509913]]AGE: 40.0 TMJOB1: 60.0 CASH: 800.0 TMADD:
60.0 INCOME: 2000.0 REGN: 3.0 LOANS: 1.0 PERS_H: 4.0
Good consumer
Probability: 【Good:Bad】
[[0.5615728 0.4384272]]AGE: 20.0 TMJOB1: 10.0 CASH: 1000.0 TMADD:
22.0 INCOME: 2000.0 REGN: 3.0 LOANS: 1.0 PERS_H: 4.0
Bad consumer
Probability: 【Good:Bad】
[[0.29095196 0.70904804]]
```

Figure 1.21 Historical predict result

2. 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.

Topic of this dissertation is ‘Assessment of the credit risk of consumer loan’. Our project is using machine learning model to reduce loan risk. In this project we need to get data, clean data, build model and develop a software to predict a consumer whether pay back on time. In this project we used python and SAS to clean and analyze data.

2.1 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.

1 – Our machine learning method in this thesis (random forest). This method is aimed to reduce risk of consumer loans. The advantage is that this algorism has high accuracy and it need not too much data cleaning.
2 – Classical method (logistic regression). It's an old method, banks used to choose it. It needs a lot of data cleaning.
3 – Classical method (linear regression). This method is easy and it has low model accuracy.

Table 2.1 Evaluation card for comparison of competitive technical solutions

Evaluation criteria ***example***	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_1	P_2	P_3	C_1	C_2	C_3
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Energy efficiency	0.1	5	5	5	0.5	0.5	0.5
2. Reliability	0.2	5	4	4	1	0.8	0.8
3. Ease of operation	0.2	5	4	4	1	0.8	0.8
4. Functional capacity	0.1	4	5	5	0.4	0.5	0.5
Economic criteria for performance evaluation							
1. Development cost	0.15	4	3	5	0.6	0.45	0.75
2. Market penetration rate	0.15	5	4	4	0.75	0.6	0.6
3. Expected lifecycle	0.1	4	5	3	0.4	0.5	0.3
Total	**1**	**32**	**30**	**29**	**4.65**	**4.15**	**4.25**

According to calculations, our method meets the requirements better than the competitors in terms of technical and economic criteria. Further investment for this project can be considered reasonable.

2.2 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 2.2 SWOT analysis

	Strengths: S1. Use machine learning to solve the complex task S2. Automatic calculation S3. Accurate prediction	**Weaknesses:** W1. Use a lot of data to create model W2. Binding to the Data just for each Bank.
Opportunities: O1. Has strong commercial value O2. New method to reduce risk	*Strategy which based on strengths and opportunities:* Use our method in bank to reduce financial risk with high accuracy and it is easy to use.	*Strategy which based on weaknesses and opportunities:* Get a lot of user information to create model and use it in one bank.
Threats: T1. The existence of classic method T2. The Invention of Better Machine Learning Algorithms	*Strategy which based on strengths and threats:* Make a comparation of our method and classic method. We could get our advantage, and we should focus on our advantage to solve problems on bank.	*Strategy which based on weaknesses and threats:* Use this method in a big bank with a lot of users. This bank need not only one method to solve problems.

2.3 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 2.3 Stakeholders of the project

Project stakeholders	Stakeholder expectations
The Bank which wants to reduce loan risk	Use machine learning model to decide whether loan money to consumer.

Table 2.4 Purpose and results of the project

Purpose of project:	Reduce loan risk for bank
Expected results of the project:	Create a machine learning model to help bank make a decision
Criteria for acceptance of the project result:	Model work correctly with high accuracy
Requirements for the project result:	Dataset of bank
	Python and SAS

The organizational structure of the project

It is necessary to solve these 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 2.5 Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours (working days (from table 7) × 6 hours)
1	Supervisor: Е.И. Губин	Head of project	Set goals and objectives, give advices, review master's thesis	96
2	Student: Li Ke	Executor	Analyze data, clean data, create machine learning model, create calculation program, write thesis	378

Project limitations

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

Table 2.6 Project limitations

Factors	Limitations / Assumptions
3.1. Project's budget	172171.7
3.1.1. Source of financing	TPU
3.2. Project timeline:	01/02/2022-01/06/2022
3.2.1. Date of approval of plan of project	01/02/2022
3.2.2. Completion date	20/05/2022

Project Schedule

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

Table 2.7 Project Schedule

Job title	Duration, working days Only working days without weekends and holidays	Start date	Date of completion	Participants
Set goals and objectives	8	01/02/2022	10/02/2022	Supervisor
Data cleaning	13	11/02/2022	01/03/2022	Student
Data modeling	13	02/03/2022	20/03/2022	Student
Create a program	15	21/03/2022	10/04/2022	Student
Prepare thesis	22	11/04/2022	10/05/2022	Student
Check thesis	8	11/05/2022	20/05/2022	Supervisor

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 2.8 A Gantt chart

№	Activities	Participants	T_c, days	Duration of the project											
				February			March			April			May		
				1	2	3	1	2	3	1	2	3	1	2	3
1	Set goals and objectives	Supervisor	8	▨											
2	Data cleaning	Student	13		■	■									

3	Data modeling	Student	13				■	■							
4	Create a program	Student	15						■	■					
5	Prepare thesis	Student	22								■	■	■		
6	Check thesis	Supervisor	8											▨	

2.4 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;

$N_{\text{cons}i}$ – 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 2.9 Material costs

Name	Amount	Price per unit, rub.	Material costs, rub.
A4 Printer paper	1	300	300
Printer toner cartridge	1	1700	1700
Total			2000

Costs of special equipment

In this project, we used computers. Supervisor's computer was provided by Tomsk polytechnic university, and student' laptop was purchase private. There is no need to buy new computers.

In this project, we used python to clean and create model. Python is free for users. Also, we used SAS (STATISTICAL ANALYSIS SYSTEM), this software needs to pay, and we used SAS free version for students.

So, we don't have costs for special equipment like computers and software.

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:

$$S_{\text{b}} = S_a \cdot T_{\text{w}}\,, \tag{3.3}$$

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:

$$S_d = \frac{S_m \cdot M}{F_{\text{v}}} , \qquad (3.4)$$

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 = 10.4 months, 6 day per week;

F_{v} – valid annual fund of working time of scientific and technical personnel (251 days).

Table 2.10 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:

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

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 2.11 Calculation of the base salaries

Performers	S_{base}, rubles	k_{reg}	S_{month}, rub.	W_d, rub.	$T_{p,}$ work days (from table 7)	$W_{base,}$ rub.
Supervisor	37700	1,3	49010	2030.7	16	32,491.2
Student	19200		24,960	1,034.2	63	65,154.6

Additional salary

This point includes the amounts 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:

$$W_{add} = k_{extra} \cdot W_{base}, \quad (x)$$

where W_{add} – additional salary, rubles;

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

W_{base} – base salary, rubles.

Total salary:

Supervisor: 32,491.2*1.1=35,740.3

Sudent: 65,154.6*1.1=71,670.1

Total additional salary:

(32,491.2+65,154.6) *0.1=9764.6

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:

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

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 2.12 Labor tax

	Supervisor	**Student**
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	35,740.3	71,670.1
Labor tax, rubles	9685.6	19422.6

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 2.13 Overhead

	Project leader	**Engineer**

Overhead rate	30%	
Salary, rubles	35,740.3	71,670.1
Overhead, rubles	10,722.1	21,501

Other direct costs

Energy costs for equipment are calculated by the formula:

$$C = P_{el} \cdot P \cdot F_{eq},$$

where P_{el} – power rates (5.8 rubles per 1 kWh);

P – power of equipment, kW;

F_{eq} – equipment usage time, hours.

C=5.8*660*0.06=230 rub

Internet: 300*4=1200 rub

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.

Table 2.14 Budget for the scientific and technical research

Name	Cost, rubles
1. Material costs	2000
2. Equipment costs	0
3. Basic salary	97,645.8
4. Additional salary	9764.6
5. Labor tax	29,108.2
6. Overhead	32,223.1
7. Other direct costs	1430
Total planned costs	172,171.7

2.5 Conclusion of financial management

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 include:

- 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.

3. Social responsibility

3.1. Introduction

The developed project aims to use data mining method and create machine learning model to analyze consumer personal information from bank dataset and create a credit risk calculator to predict the possibility of being a good consumer or a bad consumer by using Python. 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 TPU No.11 dormitory room 206. In this room has one door and one window, and there has one personal laptop with mouse and keyboard on the desk.

3.2. Legal and organizational issues of occupational safety

Nowadays one of the main ways to radical improvement of all prophylactic work referred to reduce Total Incidents Rate and occupational morbidity is the widespread implementation of an integrated Occupational Safety and Health management system. That means combining isolated activities into a single system of targeted actions at all levels and stages of the production process.

Occupational safety is a system of legislative, socio-economic, organizational, technological, hygienic and therapeutic and prophylactic measures and tools that ensure the safety, preservation of health and human performance in the work process.

The labor code of the Russian Federation states that normal working hours may not exceed 40 hours per week, The employer must keep track of the time worked by each employee [1].

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.

Since working with this library in an enterprise implies collection and analysis of personal data. To limit access to medical data and ensure their security, data processing must be carried out in accordance with federal law on the protection of personal data.

Data on the basis of the Federal Law "On Personal Data" [2]:

– The processing of personal data must be carried out on a lawful and fair basis.

– The processing of personal data should be limited to the achievement of specific, predetermined and legitimate purposes. Processing of personal data that is incompatible with the purposes is not allowed to collect personal data.

– It is not allowed to merge databases containing personal data processed for purposes incompatible between themselves.

– Only personal data that correspond to the purposes of their processing are subject to processing.

– The content and scope of the processed personal data must correspond to the stated purposes of processing. Processed personal data should not be redundant in relation to the stated purposes of their processing.

3.3. Basic ergonomic requirements for the correct location and arrangement of researcher's workplace

According to SanPiN 2.2.2 / 2.4.1340-03 [3], the workplace when working with a PC should be at least 6 square meters. The legroom should correspond to the following parameters: the legroom height is at least 600 mm, the seat distance to the lower edge of the working surface is at least 150 mm, and the seat height is 420 mm.

It is worth noting that the height of the table should depend on the growth of the operator.

The following requirements are also provided for the organization of the workplace of the PC user: The design of the working chair should ensure the maintenance of a rational working posture while working on the PC and allow the posture to be changed in order to reduce the static tension of the neck and shoulder muscles and back to prevent the development of fatigue.

The type of working chair should be selected taking into account the growth of the user, the nature and duration of work with the PC. The working chair should be lifting and swivel, adjustable in height and angle of inclination of the seat and back, as well as the distance of the back from the front edge of the seat, while the adjustment of each parameter should be independent, easy to carry out and have a secure fit.

3.4. Occupational safety

Workplace safety is the responsibility of everyone in the organization.

Occupational hygiene is a system of ensuring the health of workers in the process of labor activity, including legal, socio-economic, organizational and technical, sanitary and hygienic, treatment and prophylactic, rehabilitation and other measures.

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

Harmful production factor is a factor of the environment and the work process that can cause occupational pathology, temporary or permanent decrease in working capacity, increase the frequency of somatic and infectious diseases, and lead to impaired health of the offspring.

Hazardous production factor is a factor of the environment and the labor process that can cause injury, acute illness or sudden sharp deterioration in health, death.

In this subsection it is necessary to analyze harmful and hazardous factors that can occur during research in the laboratory, when development or operation of the designed solution (on a workplace).

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 3.1 Potential hazardous and harmful production factors

Factors (**GOST 12.0.003-2015)**	Legislation documents
1.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. Abnormal microc limatic parameters of the air	**GOST 12.1.005-88** General sanitary requirements for working zone air.
3. Increased level of noise	**GOST 12.1.003-2014** Occupational safety standards system. Noise. General safety requirements;
4. Electromagnetic fields	**SanPiN 2.2.4.1329-03** Requirements for protection of personnel from the impact of impulse electromagnetic fields.

5. Increased voltage in an electrical circuit, the closure of which can pass through the human body	**GOST 12.1.030-81** Electric safety. Protective conductive earth, neutralling.
6. Physical overload (static - long-term preservation of a certain posture)	**GOST 12.2.032-78 SSBT** Occupational safety standards system.Operator's location in a sitting position.General ergonomic requirements

Lack of natural light, insufficient illumination

The harmful effect of lighting parameters is manifested in the absence or lack of natural light, as well as insufficient illumination of the working area. Properly designed and rationally executed lighting of production facilities has a positive impact on workers, improves efficiency and safety, reduces fatigue and injuries, and also maintains high performance.

Visual discomfort and physiological strain such as anxiety, fatigue, lethargy, headaches, eyestrain, migraine, nausea, back pain, neck pain, shoulder pain, poor concentration or lack of mental alertness, and daytime sleepiness among video display terminal (VDT) workers are primarily connected with inadequate lighting in the working place and in most cases decrease work performance and efficiency.

According to the SanPiN 2.2.1/2.1.1.1278-03 standard [4], 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%.

Abnormal microclimatic parameters of the air

Computer workstations, lighting and electronic devices generate heat constantly, causing variations in the microclimate in the work environment. Increased evidence shows that indoor environmental conditions substantially influence health and productivity. High air temperatures can aggravate medical conditions and illnesses such as high blood pressure or heart disease, in addition, it causes rapid overheating of the body and accelerates fatigue. In addition, relative humidity less than 30% can lead to skin and throat irritation, producing high static.

According to the General sanitary requirements for working zone air (GOST 12.1.005-88) a computer workstation worker belongs to Category-I work, which implies light physical work. Parameters such us: air temperature, humidity and air speed determine the thermal comfort. Optimal air parameters for a Category-I work are shown in Table 2.

Table 3.2 Optimal temperature conditions for a Category I workplace

Season	Air Temperature	Humidity	Air Speed
Summer	23-26°C	~50%	<0.15m/s
Winter	20-23.5°C	~50%	<0.15m/s

In general, computer rooms must be equipped with cooling equipment (HVAC – Heat Ventilation Air Condition), which can reduce the risk of static electricity, and prevent the development of hot spots during both summer and winter.

Increased level of noise

Air-condition and equipment cooling fans necessary for the proper operation of IT equipment run continuously and create excessive noise that influence comfort, poses risk to hearing and impairs communication and concentration.

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 [5], 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.

Electromagnetic fields

In this case, the sources of increased intensity of the electromagnetic field are a personal computer. According to the SanPiN 2.2.4.1329-03 standard [6], 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.

Personal protective equipment when working on a computer includes spectral computer glasses to improve image quality and Protection against 56 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.

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.

Increased voltage in an electrical circuit, the closure of which can pass through the human body

The mechanical action of current on the body is the cause of electrical injuries. According to GOST 12.1.038-81 [7], typical types of electric injuries are burns, electric signs, skin metallization, tissue tears, dislocations of joints and bone fractures.

The following protective equipment can be used as measures to ensure the safety of working with electrical equipment:

- disconnection of voltage from live parts, on which or near to which work will be carried out, and taking measures to ensure the impossibility of applying voltage to the workplace;
- posting of posters indicating the place of work;
- electrical grounding of the housings of all installations through a neutral wire;
- coating of metal surfaces of tools with reliable insulation;
- inaccessibility of current-carrying parts of equipment (the conclusion in the case of electroporation elements, the conclusion in the body of current carrying parts)

Physical overload (static - long-term preservation of a certain posture)

Individuals who use computers and do monotonous repetitive manipulations with or without objects over long periods of time may experience discomfort or pain as a result of poor posture, improper adjustment or use of workstation components or other factors which may lead to musculoskeletal disorders (MSD). In most cases, there are relatively simple and inexpensive corrective measures which can be employed to reduce the likelihood of discomfort or injury.

Musculoskeletal conditions are typically characterized by pain (often persistent) and limitations in mobility, dexterity and overall level of functioning, reducing people’s ability to work. Low back pain is the main contributor to the overall burden of musculoskeletal conditions which are also the highest contributor to the global need for rehabilitation. Regardless of the working position, sitting for long periods of time is unhealthy.

According to the GOST 12.2.032-78 SSBT standard [8], 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.

Table 3.3 Break’s schedule when working with computers

Continuous Work	**Rest**
210 minutes	30 minutes
60 minutes	10 minutes
30 minutes	5 minutes
15 minutes	1 minute

3.5. 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 [9].

Fluorescent lamps used for artificial illumination of workplaces also require special disposal, because they contain from 10 to 70 mg of mercury, 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 [9].

3.6. Emergency Safety

In the working environment of the PC operator, according to GOST R12.1.004-85 [10], the following manufactured emergencies may occur:

-Fires and explosions in buildings and communications;

-Collapse of buildings.

Possible natural disasters include meteorological (hurricanes, showers, frosts), hydrological (floods, floods, flooding), and natural fires.

Emergencies of a biological and social nature include epidemics, epizootics, and epiphytotic. Environmental emergencies can be caused by changes in the state, lithosphere, hydrosphere, atmosphere and biosphere asa result of human activities.

The most typical for the object where the working rooms are located,equipped with a personal computer, the emergency is a fire. Premises for work of PC operators according to the classification system of categories premises for explosion and fire hazard belongs to category D (out of 5 categories A, B, B1-B4, D, D), because applies to premises with non- combustible substances and materials in a cold state [11].

All employees of the organization must be familiar with the fire safety instructions, undergo safety instructions and strictly observe it. It is forbidden to use electrical appliances in conditions that do not meet the requirements of the manufacturer's instructions, or have various kinds of malfunctions that, in accordance with the instructions for use, may lead to a fire, as well as use electrical wires and cables with damaged or lost protective properties of insulation.

Before leaving the office, it is required to inspect it, close the windows, and make sure that there are no sources of possible ignition in the room, all electrical appliances are turned off and the lighting is turned off.

With a frequency of at least once every three years, it is necessary to measure the insulation resistance of current-carrying parts of power and lighting equipment. The increase in sustainability is achieved through the implementation of appropriate organizational and technical measures, training of personnel to work in emergencies.

Upon detecting a fire or signs of combustion (smoke, burning smell, temperature increase, etc.), an employee must:

- It is required to stop work, call the fire department by phone "01";

- If possible, take measures to evacuate people and material values;
- Disconnect electrical equipment from the mains;
- Start extinguishing the fire with the available fire extinguishing means;
- Inform the immediate or superior supervisor and notify the surrounding employees;
- In case of a general signal of danger, leave the building in accordance with the "Plan for the evacuation of people in case of fire and other emergencies."

To extinguish a fire, use manual carbon dioxide fire extinguishers (type OU-2, OU-5) located in the office premises, and a fire hydrant internal fire-fighting water supply. They are designed to extinguish the initial fires of various substances and materials, with the exception of substances that burn without air access. Fire extinguishers must be kept in good working order at all times and ready for action. It is strictly forbidden to extinguish fires in office premises using chemical foam fire extinguishers (type OHP-10) [12].

3.7. Conclusion of social responsibility

Each employee must carry out professional activities with taking into account social, legal, environmental and cultural aspects, issues health and safety, be socially responsible for the solutions, be aware of the need for sustainable development.

In presently section covered the main issues of observance of rights employee to work, compliance with the rules for labor safety, industrial safety, ecology and resource conservation.

It was found that the researcher's workplace satisfies safety and health requirements during project implementation, and the harmful impact of the research object on the environment is not exceeds the norm.

3.8. Reference of social responsibility

1.Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018).

2. Федеральный закон «О персональных данных». Книга 2, ст. № 5. — 2006. — С. 28.

3. SanPiN 2.2.2 / 2.4.1340-03. Sanitary-epidemiological rules and standards "Hygienic requirements for PC and work organization".

4. SanPiN 2.2.1/2.1.1.1278-03. Hygienic requirements for natural,artificial and mixed lighting of residential and public buildings.

5. GOST 12.1.003-2014. Occupational safety standards system. Noise. General safety requirements;

6. SanPiN 2.2.4.1329-03. Requirement for protection of personnel from the impact of impulse electromagnetic fields.

7. GOST 12.1.030-81 Electric safety. Protective conductive earth, neutralling.

8. GOST 12.2.032-78 SSBT. Occupational safety standards system.Operator's location in a sitting position.General ergonomic requirements.

9. GOST R ISO 1410-2010. Environmental management. Assessment of life Cycle. Principles and structure.

10. GOST R12.1.004-85 Occupational safety standards system. Fire safety

11. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements.

12. GOST Industrial equipment. General safety requirements to working places.

Conclusion

In this paper, we used SAS and Python assess the credit risk of consumer loan, including data description, data preparation, creation of model, model accuracy test, feature selection and develop a graphic user interface - credit risk calculator. We used random forest algorithm and logistic regression to create model, both them are good model on solve binary classification problem. We got 8 important features from model, and used these features to create a calculator to help banks and other financial institutions control the risk of borrowers who want to loan money. Also, this calculator let you know which features consumers need to take care of when he wants to loan money from bank.

Reference

1. Advantages of SAS in data analysis, URL: https://zhuanlan.zhihu.com/p/62487120
2. Why should we use python for data analysis? URL: https://zhuanlan.zhihu.com/p/48226550
3. SEMMA-Wikipedia, URL:https://en.wikipedia.org/wiki/SEMMA
4. Data analysis-SEMMA Steps, URL: https://blog.csdn.net/zfh_0916/article/details/106673553
5. Таворнпрадит П. Developing credit risk score using SAS programming: магистерская диссертация / П. Таворнпрадит ; Национальный исследовательский Томский политехнический университет (ТПУ), Инженерная школа информационных технологий и робототехники (ИШИТР), Отделение информационных технологий (ОИТ) ; науч. рук. Е. И. Губин. — Томск, 2021.
6. Губин Е. И. Методология подготовки больших данных для прогнозного анализа / Е. И. Губин // Современные технологии, экономика и образование : сборник трудов Всероссийской научнометодической конференции, г. Томск, 27-29 декабря 2019 г. — Томск : Изд-во ТПУ, 2019. — [С. 27-29].
7. Introduction of random forest algorithm, URL: https://www.jianshu.com/p/ae1826eb7836
8. Random forest, URL:https://easyai.tech/ai-definition/random-forest/
9. ROC and AUC, URL: https://zhuanlan.zhihu.com/p/58587448

Appendix A. Program code for credit risk calculator (GUI)

```
import sys
import os
import numpy as np
import pandas as pd
from PyQt5.QtWidgets import QApplication, QWidget, QPushButton, QLabel,
QTableWidget, QTableWidgetItem, QPlainTextEdit, \
    QVBoxLayout, QTextBrowser,QHBoxLayout,QLineEdit
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import make_pipeline
from sklearn.ensemble import RandomForestRegressor
from PyQt5 import QtCore
from PyQt5.QtCore import QCoreApplication
import joblib
model1 = joblib.load("train_model.m")

app = QApplication(sys.argv)
window = QWidget()
window.resize(1000, 200)
window.move(560, 350)
window.setWindowTitle("Credit Risk Calculator")

def rf():
    text.clear()
    t1 = QLineEdit1.text()
    t2 = QLineEdit2.text()
    t3 = QLineEdit3.text()
    t4 = QLineEdit4.text()
    t5 = QLineEdit5.text()
    t6 = QLineEdit6.text()
    t7 = QLineEdit7.text()
    t8 = QLineEdit8.text()
    t1 = float(t1)
    t2 = float(t2)
    t3 = float(t3)
    t4 = float(t4)
    t5 = float(t5)
    t6 = float(t6)
    t7 = float(t7)
    t8 = float(t8)
    test=np.array([t1,t2,t3,t4,t5,t6,t7,t8]).reshape(1, -1)
    result=model1.predict(test)
    result=str(result)
    result1=model1.predict_proba(test)
    result1=str(result1)

    if result =='[1]':
        text.append('AGE：'+str(t1)+' '+'TMJOB1: '+str(t2)+' '+'CASH:
'+str(t3)+' '+'TMADD: '+str(t4)+' '+'INCOME: '+str(t5)+' '+'REGN:
'+str(t6)+' '+'LOANS: '+str(t7)+' '+'PERS_H: '+str(t8))
        text.append('Bad consumer')
        text.append('Probability：【Good:Bad】')

        text.append(result1)
        aaa=text.toPlainText()
        f = open('result.txt',mode='a+')
        f.write(aaa)
        f.close()
    else:
        text.append('AGE：'+str(t1)+' '+'TMJOB1: '+str(t2)+' '+'CASH:
'+str(t3)+' '+'TMADD: '+str(t4)+' '+'INCOME: '+str(t5)+' '+'REGN:
'+str(t6)+' '+'LOANS: '+str(t7)+' '+'PERS_H: '+str(t8))
```

```
        text.append('Good consumer')
        text.append('Probability：【Good:Bad】')
        text.append(result1)
        aaa=text.toPlainText()
        f = open('result.txt',mode='a+')
        f.write(aaa)
        f.close()
    return
def sv():

    os.system('result.txt')
    return

QLineEdit1 = QLineEdit(window)
QLineEdit1.setPlaceholderText('AGE')
QLineEdit2 = QLineEdit(window)
QLineEdit2.setPlaceholderText('TMJOB1')
QLineEdit3 = QLineEdit(window)
QLineEdit3.setPlaceholderText('CASH')
QLineEdit4 = QLineEdit(window)
QLineEdit4.setPlaceholderText('TMADD')
QLineEdit5 = QLineEdit(window)
QLineEdit5.setPlaceholderText('INCOME')
QLineEdit6 = QLineEdit(window)
QLineEdit6.setPlaceholderText('REGN')
QLineEdit7 = QLineEdit(window)
QLineEdit7.setPlaceholderText('LOANS')
QLineEdit8 = QLineEdit(window)
QLineEdit8.setPlaceholderText('PERS_H')

button1 = QPushButton('Predict', window)
button1.clicked.connect(rf)
```

```
button2 = QPushButton('Log', window)
button2.clicked.connect(sv)

text = QTextBrowser(window)

layout1 = QHBoxLayout()
layout1.addWidget(QLineEdit1)
layout1.addWidget(QLineEdit2)
layout1.addWidget(QLineEdit3)
layout1.addWidget(QLineEdit4)
layout1.addWidget(QLineEdit5)
layout1.addWidget(QLineEdit6)
layout1.addWidget(QLineEdit7)
layout1.addWidget(QLineEdit8)

layout = QVBoxLayout()
layout.addLayout(layout1,1)
layout.addWidget(button1,1)
layout.addWidget(text,1)
layout.addWidget(button2,1)

window.setLayout(layout)

window.show()
sys.exit(app.exec_())
```