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
Developing credit risk score using SAS programming

УДК 004.438:004.45:336.774

Студент

Группа	ФИО	Подпись	Дата
8ПМ9И	Таворнпрадит Пхана		21.06.2021 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		21.06.2021 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Гончарова Н. А.	к.э.н.		22.02.2021 г

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		19.02.2021 г.

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		21.06.2021 г

Томск – 2021

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**
8ПМ9И	Таворнпрадит Пхана

Тема работы:

Developing credit risk score using SAS programming	
Утверждена приказом директора (дата, номер)	№ 40-5/с от 09.02.2021

Срок сдачи студентом выполненной работы:	15.06.2021

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 SAS 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.э.н., доцент Гончарова Н. А.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2021

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н		1.03.2021

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ9И	Таворнпрадит Пхана		1.03.2021

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 2020 /2021 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:	28.06.2021

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
01.06.2021	Основная часть	70
01.06.2021	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
01.06.2021	Социальная ответственность	10
01.06.2021	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
8PM9I	Tavornpradit Phana

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 of the scientific supervisor - 46,284.34 rubles. - Student scholarship - 12,900.16 rubles.
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs – 1371.99
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Additional salary – 13977.87 – Overhead costs – 8136.3
The list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	– Use machine learning and doing automation calculation
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	N.A. Goncharova	PhD		22.02.2021

The task was accepted by the student:

Group	Full name	Signature	Date
8PM9I	Tavornpradit Phana		22.02.2021

Task for section

«Social responsibility»

To student:

Group	Full name		
8PM9I	Tavornpradit Phana		
School	Information Technology and Robotics	**Department**	Information Technology
Degree	Master programmer	**Specialization**	09.04.04 Software Engineering

Title of graduation thesis:

Developing credit risk score using SAS programming	
Initial data for section «Social Responsibility»:	
1. Information about object of investigation (matter, material, device, algorithm, procedure, workplace) and areaof its application	– Base exploratory data analysis to make the bank easier to make decision – Machine learning algorithm to select most important variables in borrower's profiles – Working area: desktop and personal computer
List of items to be investigated and to be developed:	
1. Legal and organizational issues to provide safety: – Special (specific for operation of objects of investigation, designed workplace) legal rules of labor legislation; – Organizational activities for layout of workplace.	– 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. Work Safety: 2.1. Analysis of identified harmful and dangerous factors 2.2. Justification of measures to reduce probability ofharmful and dangerous factors	Dangerous and harmful factors: – Increased levels of electromagnetic radiation; – Insufficient illumination of workplace – Excessive noise – Increased / decreased air humidity in the workplace; – Electric shock – Ionizing radiation
3. Ecological safety:	–Insufficiency of atmosphere and hydrosphere –Lithosphere: when disposing of fluorescent lamps and office equipment
4. Safety in emergency situations:	– Fire safety
Assignment date for section according to schedule	

The task was issued by consultant:

Position	Full name	Scientific degree, rank	Signature	date
docent professor	Antonevich O.A	PhD		19.02.2021

The task was accepted by student

Group	Full name	Signature	date
8PM9I	Tavornpradit Phana		19.02.2021

Abstract

Final qualifying work 83 pages, 14 figures, 23 tables, 20 sources.

Keywords: data analysis, data preparation, data cleaning, logistic regression, credit scoring, credit rating, scorecard, credit risk model.

The object of the study is data on the creditworthiness of borrowers.

The subject of the research the credit risk model and make a system interface (GUI).

The objective of the project is to develop the system from credit risk model by using SAS programming to help the bank in decision-making, control the risk and choose more good borrowers and delete bad borrowers from the bank.

In the research project, to learn how to develop credit scoring using SAS programming and create program calculate credit risk score using Python .

As a result of the study, it is show that to improves the accuracy of the assessment of creditworthiness by using logistic regression (selection variables method, evaluation dataset). The conclusion is the program calculator is made.

Basic design, technological and technical and operational characteristics: the developed methodology allows assessing the creditworthiness of a potential borrower.

Scope: the developed technique can be used to improve the accuracy of credit scoring in banks and the calculator interface make the credit risk model more easily to use.

Publications:

XVIII Международной научно-практической конференции студентов, аспирантов и молодых ученых.

МОЛОДЕЖЬ И СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ МСИТ-2021

Томск, 22 -26 марта 2021г.

Tavornpradit Phana Natthasin (Томский политехнический университет) “The most important variables for credit risk model”

Table Content

Introduction

Banks and finance organizations very often use credit risk models for evaluating potential customers when providing credit. In the past credit risk uses way, which weights various factors including credit history, length of credit history, types of credit used, your current credit, and so on. Credit scoring is the method that will aids banks in making a decision for borrowers. This technique describes which person should get a credit score, how much credit score they should get, and which operational strategies will help the lenders know the probability to get profit from borrowers and easy to control the risk of the individual borrower.

Credit Scoring for SAS programming have to do in the following step: calculate a risk score for borrowing's application or an existing credit borrower's account, check the accuracy of the model, and then monitoring the result of the credit score that how credit risk score effect to the decision-making that have on key business performance indicators.

Although, credit scoring is not as good as pricing different financial derivatives by using a statistical model. However, this model is the successful one for the applications of the research of statistic model in terms of decision-making for finance and banking. To training data in the credit scoring, for example in this paper is real personal data of borrowers that have been anonymized for obvious reasons. The features - characteristics credit scoring - include your age, salary, people in your household, number of your children, time that you have been working in your current job, etc. For this project, the target variable is a binary variable of the ”bad” and ”good” values with respect to the borrowers defaulting given some historical period.

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

- Data preparation
- Variable selection
- WOE (weight of evidence) and IV (information value)
- Scorecard scaling
- Create user interface (Calculator)

However, this Credit Risk Model will help the bank in term of decision-making for the new borrowers, control the risk, delete bad borrowers from the bank and choose more good borrowers for the bank. This work also let us know how good borrower's profiles and bad borrower's profiles look like.

1 Object and research methods

1.1 Description of the method for constructing a scorecard using a logistic regression model.

The most common method for assessing a borrower's creditworthiness in banks is credit scoring. Credit scoring (application scoring) is an automated system based on a predictive mathematical model that uses a bank's credit history to predict the likelihood that a potential borrower will repay the loan on time [1-2]. The forecast is based on information about the credit history, social and demographic parameters, data on the requested loan. Currently, banks are paying particular attention to the analysis of credit risks due to the increased incidence of loan defaults and fraud. When building a scoring model, it is required not only to determine, in general on the basis of the assigned score, whether it is worth giving the borrower a loan or not, but also to determine the minimum score for issuing a loan.

Most banks create scoring models on their own, using their own data collected over previous years, or use ready-made solutions based on generalized data on borrowers of several banks. In both cases, the methods of building the models are trade secrets.

The most popular predictive model for building scorecards or credit risk model is the logistic regression model. From [7-9] in the paper tell that the logistic regression model can access to likelihood value of a loan payment for each borrower and the target variable of dataset in this project is in a binary format (0,1), so the logistic regression model can tell the relationship between dependence variables and independence variables.

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,1} + \ldots + + b_k x_{i,j} + \varepsilon_i \qquad (1)$$

Where p_i— the probability that the i borrower will default the loan

x_{ij} — the value of the j independent variable

b_0 — independent constant of the model, bj - model parameters

ε_i — a component of a random error

Equation (1) show how linearly dependent of the probability of a loan delinquency, depending on the values of independent variables.

1.2 Problem statement

It is required to prepare data[3-6] for solving the problem of binary classification of potential bank borrowers using the logistic regression method PROC LOGISTIC (SAS)[7-9]. The initial data for the work are historical data on creditworthiness, containing 24 variables, one of which is the target, and 3000 observations. This sample is balanced by the target variable, that is, the number of payers and Good borrowers. Bad borrowers are those borrowers who have not made the planned loan payments within 90 days. For easy to use the scorecard, It is required a calculator and interface to calculate the new borrower's information.

Table 1 shows the variables characterizing the borrowers.

Variable name	Defination	Variable type
TITLE	The nature of homeownership	Categorical
CHILDREN	Amount 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

The target variable is GB, which is 0 is mean the good borrower and 1 is mean the bad borrower who not pay their loan in 90 days..

1.3 Selection and description of the method for solving the problem

From the results of the latest KDnuggets polls (2014), 43% of respondents use the CRISP-DM data analysis methodology, 8.5% - the SEMMA methodology, 3.5% - the organization's own methodology, 27.5% - their own methodology, others

methodologies are used by 17.5% of the respondents. 0% of the respondents do not use any methodology [10].

The two leading methodologies are generally very similar, but CRISP-DM has gained popularity as being more complete and detailed than SEMMA. Each of these methodologies includes a data preparation stage, which in both cases has a rather general recommendatory nature, which leads to the need to create a clearer and more detailed methodology for preparing data for classification using logistic regression.

1.3.1 CRISP-DM Methodology

CRISP-DM is a standard cross-industry Data Mining process consisting of six stages organized in a cycle. It is currently the most popular methodology. According to the CRISP-DM standard, it includes the following steps [10-12]:

1. Business analysis (Business understanding) - the initial phase, at which the definition of business goals and requirements for results are developed.
2. Data understanding. The second phase starts with collecting data, describing, examining and checking the quality of the data.
3. Data preparation. The data preparation phase includes sampling, feature engineering, data cleaning, integration, and formatting.
4. 4. Modeling. At the modeling stage, the selection, training and quality assessment of the models are carried out.
5. Evaluation of results includes an assessment of the process, the results obtained and the definition of subsequent actions.
6. Deployment. This step involves implementing the model, monitoring and receiving feedback.

Figure 1Method CRISP-DM

1.3.2SEMMA Methodology

The SEMMA methodology, created by the SAS Institute, is an alternative to CRISP-DM. SEMMA consists of five stages [10-12]:

1. Sample - the formation of an initial dataset for modeling (dataset), which must be large enough to contain sufficient information for extraction, and it is also have limited in order to use effectively used.
2. Explore - identifying associations, visual and interactive statistical analysis, understanding data by detecting expected and unexpected relationships between variables, as well as deviations using data visualization.
3. Change (Modify) - application of methods of selection, creation and transformation of variables in preparation for modeling: cluster analysis, transformation, filtering and substitution of information.Моделирование
4. Model - application of methods for constructing and processing data mining models: neural networks, decision trees, regression analysis, etc.
5. Assess - comparison of simulation results with planned indicators, analysis of the reliability and usefulness of the created models.

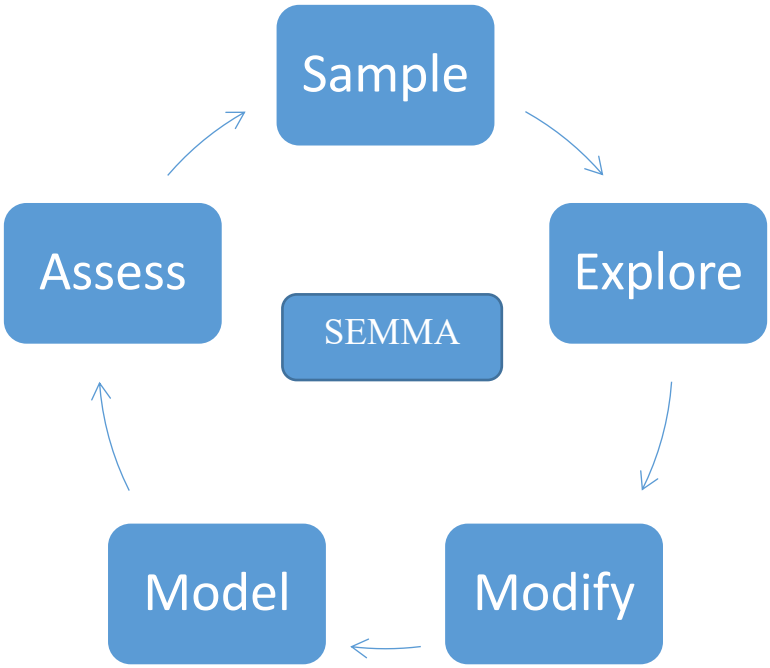

Figure 2 Method SEMMA

This methodology was developed for use in the SAS Enterprise Miner environment, so its steps are focused on the capabilities of this software package.

The SEMMA methodology is more of a set of guidelines than a set of hard rules and is less detailed than CRISP-DM. This explains it is less popularity.

1.3.3Developed data preparation methodology

1. Partitioning data

To build a credit risk model must split the data before applying it into training dataset and validation dataset before putting it into regression model.

For training prediction models, it is recommended to split the dataset into two datasets – training dataset and validation dataset. The regression model will train on the training dataset for optimize the variable and get the information and for test the result of the model we will test it on the validation dataset. The training process is so important more than the test dataset, so we should split it around 70%-80% of the data for the training dataset and 20%-30% for the validation dataset. This is depending on the amount of data that you have. A model trained on a small amount of test data has a large variance, i.e., its results on different dataset. With an average sample size (1,000 to 10,000 observations), the rate of split data 70:30 and 80:20 is the standard for the data partition of the dataset before applying it into the regression model. If your observation is exceeding 10,000 observations, you can reduce the training dataset and increase the validation dataset instead, especially when the predictive require a lot of data of computation. With a small amount of initial data (100 to 1,000 observations), it is worth to try using the cross-validation method to help [3-6].

Cross-validation is a method of dividing the initial data into training and test dataset in conditions of not have enough data. Cross-validation can reduce the variance of the model. In the process of cross-validation, the data is first split into training dataset and validation dataset. The training dataset is divided into k parts. Training takes place in k-1 parts, the rest is used for validation [3-6].

When training may models at once, you can split the data into 3 parts training dataset, test dataset and validation dataset. This is the solution for training -obtaining an optimistically biased estimate of the model. You can compare the result of many models from the validation dataset and will choose the best result out of it. The standard size of split the data in this situation in 60:20:20.The recommendation from this paragraph is if the size of the two samples are applicable. [3-6].

The above data partitioning schemes are shown in Figure 3.

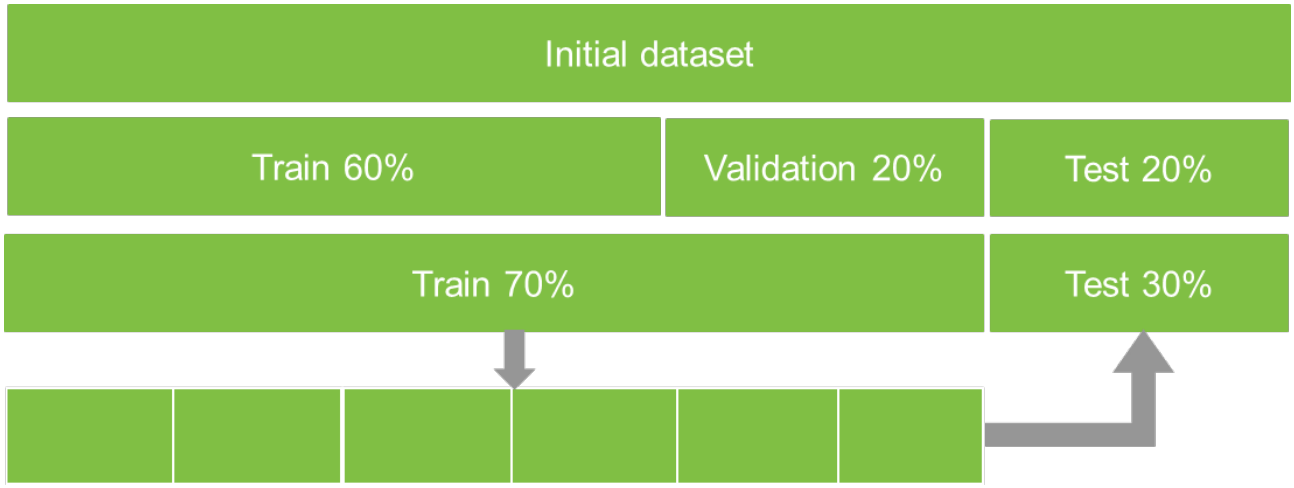

Figure 3 Partition data

When you want to split the dataset. The dataset must have the same target variable.

2. Data cleansing

Cleansing data method [3-6] includes removing duplicate records, correcting erroneous and inconsistent data, handling outliers and missing values. Solving these problems can significantly improve the quality of the predictive model.

Remove the duplicate record

The duplicate record will appear in the data collection method or especially when your dataset come from multiple datasets. This problem will lead to more problems when you want to analyze it. To remove the duplicate record from your dataset will make the analysis more better and decrease variance from the target variable and it will increase accuracy of the dataset. [3-6].

Outliers

The data that look abnormal and stand out from the other data is call “Outliers”, To manage it depends on the analysis to tell that which record is abnormal or stand out from the others.

The graph is a good way to make you easily find the outliers. You can use chart or plot to see the outliers clearly. You have to handle the outliers because it can make the result of analysis bias or not accurate.

Sometimes the outliers of the dataset is come from the result of error of input data or error in the data collection process, so it has to be removed but sometimes the outlier can also be useful in the case of you want to analyze the best case or the worse case of your dataset, these value will be always stand out from the others.

In case of mistake when input the data or in the data collection process. You can use cluster algorithm to the intention or behavior or that outliers instead of delete it and maybe you can find the best case out the worse out of it. [3-6].

Correction of inconsistent data

Incorrect and inconsistent values may appear at the stage of input, transmission and collection of data as a result of typos, software restrictions (limitation on the length of a variable, limitation of the buffer size), various data recording formats.

Rare categorical variables, extreme (height: 250 cm) or unusual (salary: -1 rubles) values of numeric variables may turn out to be incorrect. You can identify such values using histograms, a box with a mustache, or scatterplots. Conflicting data (gender: male, pregnancy: yes) can be identified using the relevant logic rules. To identify some erroneous and conflicting data, you may need a subject matter expert [3-6].

Logistic regression also can make a data inconsistent or incorrect due to the mistake when we input the data to the model. The value that causes by the logistic regression also have to be handle or remove to make the analysis more accurate.

Processing passes

Gaps in data can be due to many reasons: the necessary data may not always be available (customer information), data may be missing because it was considered unnecessary at a certain point in time. Gaps can also occur due to technical issues. Data may be deleted due to inconsistency. Many methods of data analysis, including logistic regression, are not able to work with missing data, so the gaps must be eliminated in one way or another: delete cases containing gaps or fill them in.

If it is necessary to fill in the field at the stage of data entry, the missing values are encoded with some substitute value, chosen so that it does not resemble a typical

value for the variable. Typical replacement values for data gaps are presented in Table 2.

Table 2 Typical values replacing missing data

Code	Description
Null, empty, line, “”	for a numeric or categorical variable
0	a numeric variable whose value is never zero
-1	a numeric variable that only takes positive values
99, 999	a numeric variable whose values can be less than 100, 1000, etc.
-99, -999	numeric variable that can take negative values
U, UU	categorical variable
000000, XXXXXX	postcode
11.11.11	date
00000000000	phone number

Have a lot of reasons that can generate a missing value in the dataset. The missing data will be the cause of many problems, for example it will be failure to collect the data, or the dataset will decrease its accuracy and also many machines learning models do not support the missing value, so we should solve this missing value problems seriously.

The ways to solve the missing value problem are:

1. Delete the record with the missing value, so you delete the column that have null values and if more than 50% of the column is null you can also drop it. This way will increase the accuracy of the dataset, but your information will be loss.
2. Replace the missing value with median or mean values, if the record in your dataset has a numeric continuous value, you can consider replace it with the mean or median. This way is better than delete the record in terms of loss of the data. This way will be easy to implement and good for the small dataset, but this way will work just only the numerical variable and can make the data leakage.
3. Imputation method for categorical variables (string or numerical). You can replace the missing value with the most frequency of each variable. This way your data will not loss, but it will work just only with the categorical variable.
4. Predict the missing value, in this technic, you can use logistic regression model, SVM, etc., for the categorical variable and use linear regression model, SVR, etc., for the numerical variable.

3. Transformation

Sampling

To sampling the continuous variable will be better in terms of distribution rate of the variable has the outliers or missing values. In such cases, a sampled version of the continuous can make complex nonlinear relationships easier to analyze.

The numerical variable discretization algorithm consists of three steps. First, the values of a numeric variable are divided into several groups by quantiles. For each group, the weight of the categories Weight of Evidence (WOE) is calculated using the formula.

$$WOE_i = \ln\left(\frac{d_{i1}}{d_{i2}}\right),$$

where d_{i1}, d_{i2} – relative frequencies of bad and good borrowers in the i group of the sampled variable; $i=1,\ldots,k$,

k – the number of categories of the variable.

For calculate the WOE value have 2 step, the first is to split the data into a few group (and in both cases you have to be sure that all observations in one group have the same effect on target variable),the second is to calculate the WOE value for each group. In the project [20] recommended that for each group of WOE must have at least 5% of all the value in each variable for example our target variable is GB (good/bad) for each group must have the value of good and bad together[1-2]

Normalization (scaling)

Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization. Otherwise, you can run your logistic regression without any standardization treatment on the features.

Standardization is the popular method to scaling and it will use when the variable's distribution is close to normal, otherwise min-mix normalization is preferable. Scaling methods are presented in table 3.

Table 3 Scaling methods

Scaling method	Formula	Range

Min-max normalization	$\frac{x - x_{\min}}{x_{\max} - x_{\min}}$	[0, 1]
Standardization	$\frac{x - \mu_x}{\sigma_x}$	in most cases [-3, 3]

Nonlinear transformation

Normally, the conversion method will convert all the variable values into zero or negative values. This method aids all independent variable to have maximize their relationship with target variable and the result will make the independent variables describes the target variable better. This method usually uses for convert numerical variable type for example the reciprocal, cubic, square, square root, decimal, exponential, etc. [1-2].

4. Choossing variables

For the model to improve their accuracy before analysis. The model have to pass the step of variable selection using proc logistic[7-9]. The variable that have p value >0.05 will be selected from the proc logistic[7-9] and the variable that have p value < 0.05 will be deleted to improve the model.

Multicollinearity

The variance in the logistic regression will increase when you found the multicollinearity, you will get the wrong value when estimating the parameters of the model and also inconstancy when estimating the parameters of the model.

You can use the correlation matrix and the Variance Inflation Factor (VIF) to find the multicollinearity.:

$$VIF = \frac{1}{1 - R_i^2},$$

where R_i – the coefficient of the regression of the i variable on the remaining explanatory variables.

If the VIF is more than five, this indicates the presence of multicollinearity. in the case like this, you have to delete this variable from the dataset because it will effect the analysis.

Informativeness

The main idea of the informativeness is to find, how strong of the relationship between independent variable and target variable.

You should that the variable that have no relation with the target variable like id, name, date, etc. You should delete it out of the analysis.

The main methods for assessing the informativeness of variables are the chi-square test and the indicator of information value (IV).

The IV is calculated using the following formula:

$$IV = \sum (\% \text{ of non-events} - \% \text{ of events}) * WOE$$

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Figure 4 Rule relate to information value(IV)

According to Siddiqi (2006), by convention the values of the IV statistic in credit scoring can be interpreted as follows.

1. If the IV statistic <0.02, then the predictor is useless for modeling, like the value is not clear enough to separate the good and bad
2. 0.02 >=IV<= 0.1, then the predictor not much effect to the good/bad variables
3. 0.1 >=IV<= 0.3, then the predictor has a medium effect to the good/bad variables.
4. 0.3 >=IV<= 0.5, then the predictor has a strong effect to the good/bad variables.
5. IV > 0.5, It is mean that this value is a suspicious relationship, so you should check it.

Information value is not a suitable selection variable when you want to built the classification model but for the logistic regression model, so if you use information value to select the variable in logistic regression model it will be well suitable for it

and this information value also can tell hoe confusion of the predictive value be. You can say that this information value is designed for binary logistic regression model.

1.4 Selection and description of the software environment

Three tools were chosen to implement the data preparation methodology: Python 3 in the Jupyter Notebook shell and SAS.

Python – a programming language with a relatively low threshold of entry, currently the most popular language for data analysis, has many open libraries.

Jupyter Notebook the program code is located in a series of cells - executable or markup. Layout cells support LaTeX, which allows you to use mathematical expressions in them. Files generated in Jupyter Notebook are in .ipynb format, which is equivalent to .json format. The resulting files can be stored in a version control system.

In this project, the following libraries were used:

1. NumPy (Numerical Python) – provides optimized functions for working with multidimensional data arrays.
2. Pandas (Panel Data) – to manage the data into data panel to easy to manage.
3. tkinter – Building the GUI interface.

SAS (Statistical Analysis System) – software package for data processing and analysis, the market leader in business analysis

SAS Studio has powerful features that let you efficiently prepare and orchestrate your data for better decisions.

In this project, the following libraries were used:

1. proc logistic – to use logistic regression model.
2. proc transreg – Create the new output data.
3. proc means- used to check the basic statistical analysis of the variables in the dataset.

4. proc rank – to group and indicate the rang of the variable(bin)
5. proc standard – check the standard deviation value.
6. proc univariate – help to check the Outliers of the dataset by show the distribution rate.
7. proc freq – check the frequency of the variable in the dataset
8. proc format – Create the string format for checking the format of the string.
9. proc sql – can manage the data using the sql context.
10. proc Surveyselect – help to split the data into training dataset and the validation dataset .

2 Software implementation of the technique

The algorithm of the proposed methodology is implemented in three software environments.

All software implementations read the source data from the file in the .xls format.

- The Python implementation is done using the Jupyter Notebook web shell. Consists of one .ipynb file.
- A SAS implementation consists of several .sas files containing the implementation of individual functions.

3 Results of the study

Consider the results of applying the proposed methodology to data on borrowers.

3.1 Splitting the original sample

Since the amount of initial data is small (3000 observations) and only one predictive model is used, the initial sample is divided into two - training and test in a ratio of 70% to 30% by proc Surveyselect.

3.2 Data cleansing

Removing duplicate lines

There are no duplicate rows in this sample.

Outliers

Analysis of univariate outliers in numeric variables using range diagrams (Fig. 5) and histograms (Fig. 6) shows that the variables CHILDREN, PERS_H, INCOME, CASH contain values that can be an error or outlier in the data. Variables LOANS and AGE contain outliers, while in variables TMADD and TMJOB1 the missing values are encoded with 999. Extremely large values of CASH and INCOME are most likely encoded gaps in the data, since they do not agree with other data, as shown in Figure 3.4. The values are in the variables TMADD, TMJOB1, CASH and INCOME in the amount of 93, 34 11 and 1 pc. respectively were replaced by passes. Also, an observation with extreme values for the variables CHILDREN, PERS_H was removed from the analysis, which are consistent with each other (23 children and 25 living in the house), and, therefore, are an outlier, not an error or a skipping code,

but can strongly affect the regression coefficients

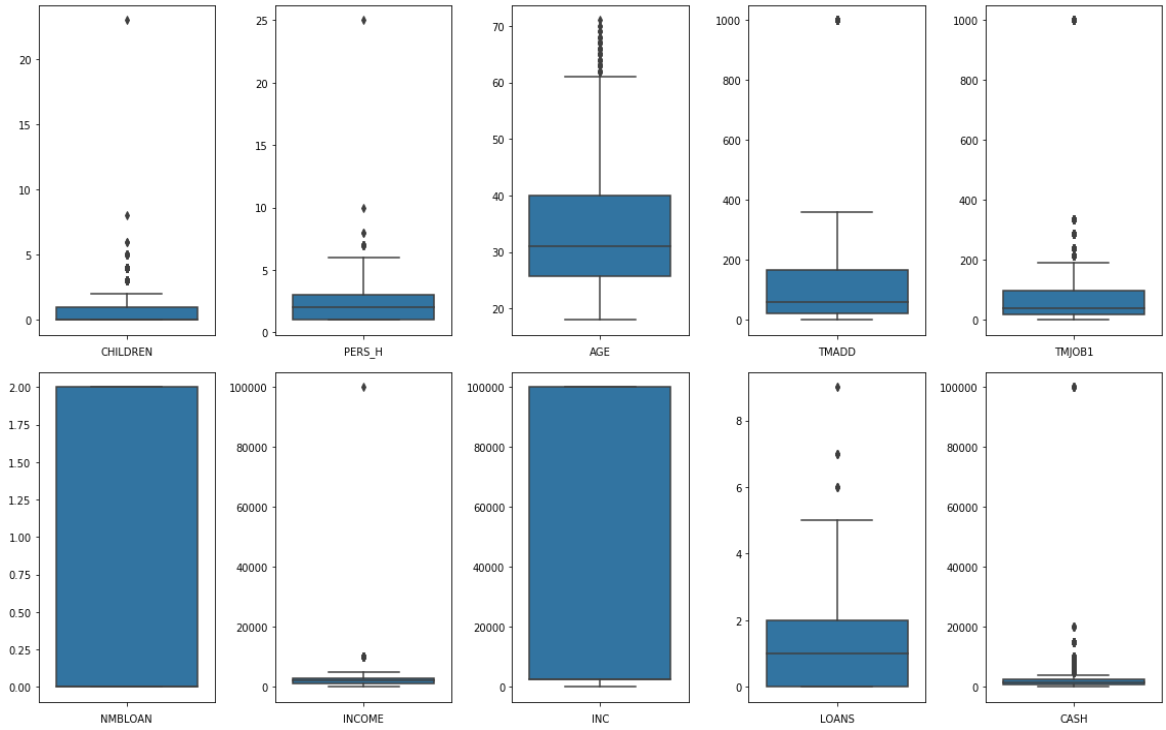

Figure 5 Box plot

In Fig.6 You can assumed that none of the explanatory variables is normally distributed.

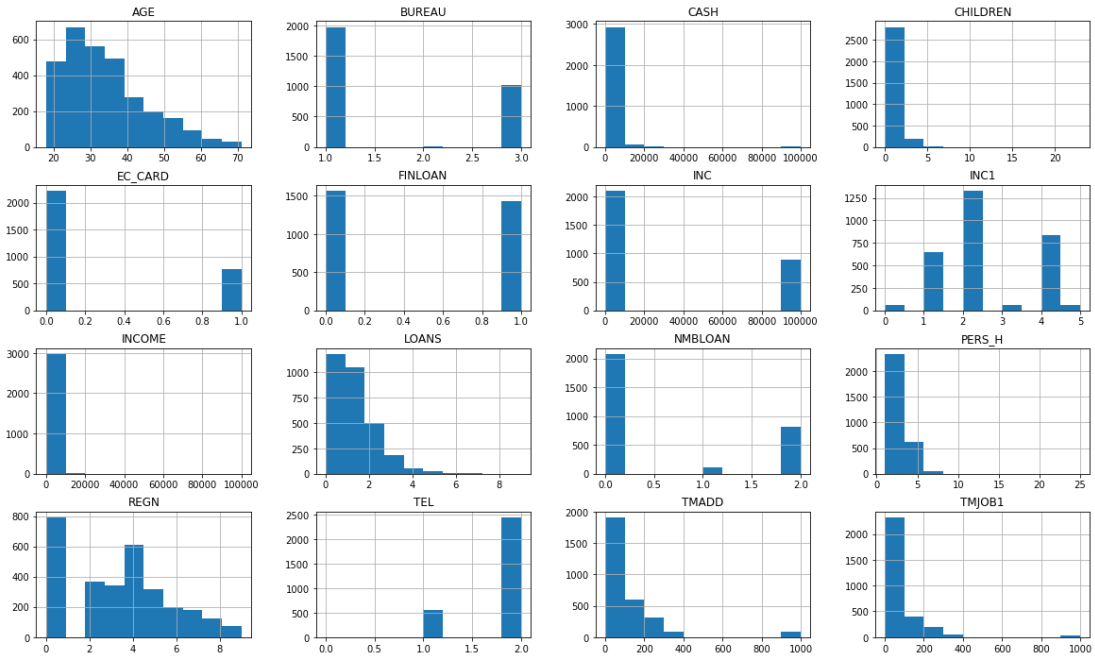

Figure 6 Raw data histograms

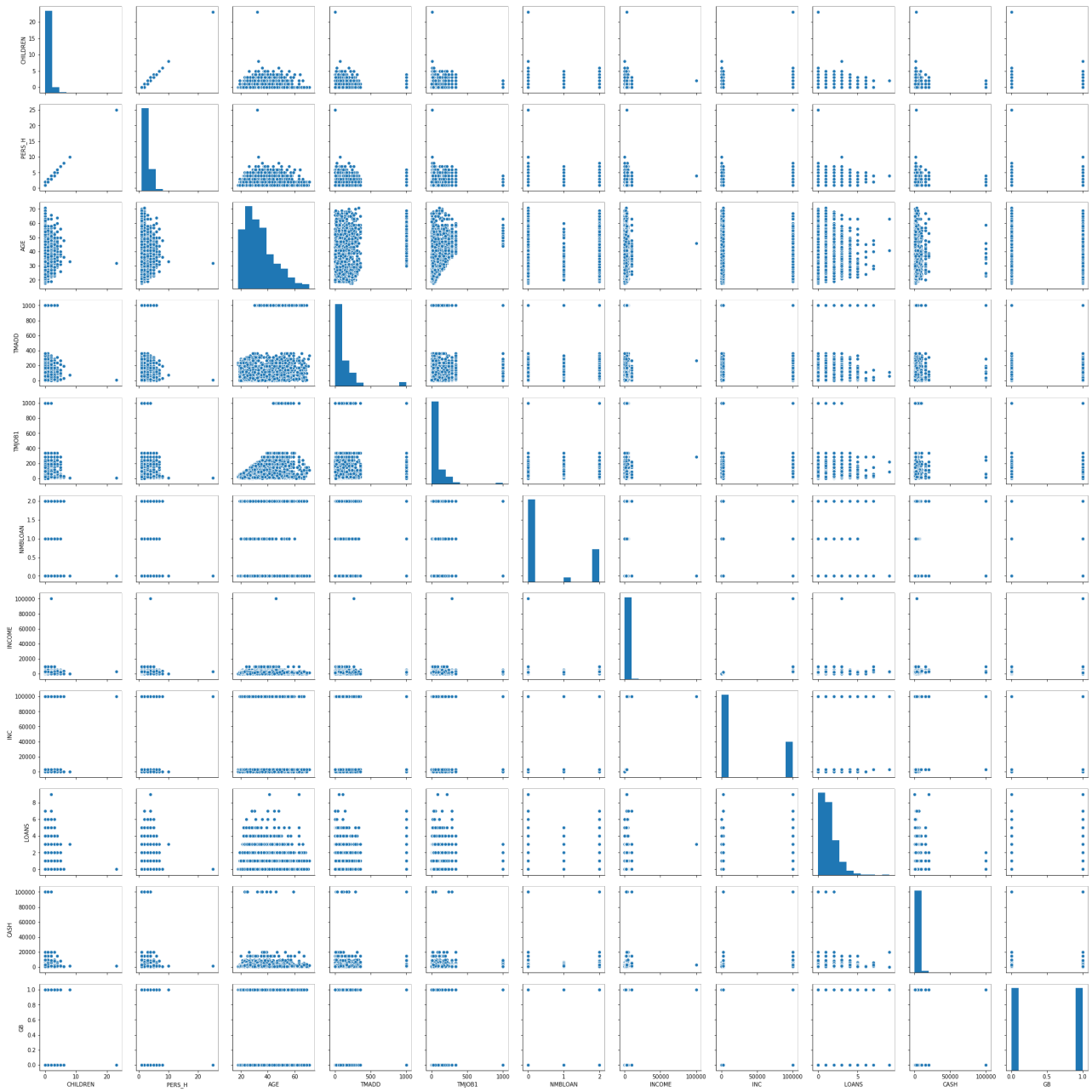

Figure 7 Scatter diagrams of raw data

In Fig. 8 shows the peak-to-peak diagrams after outliers.

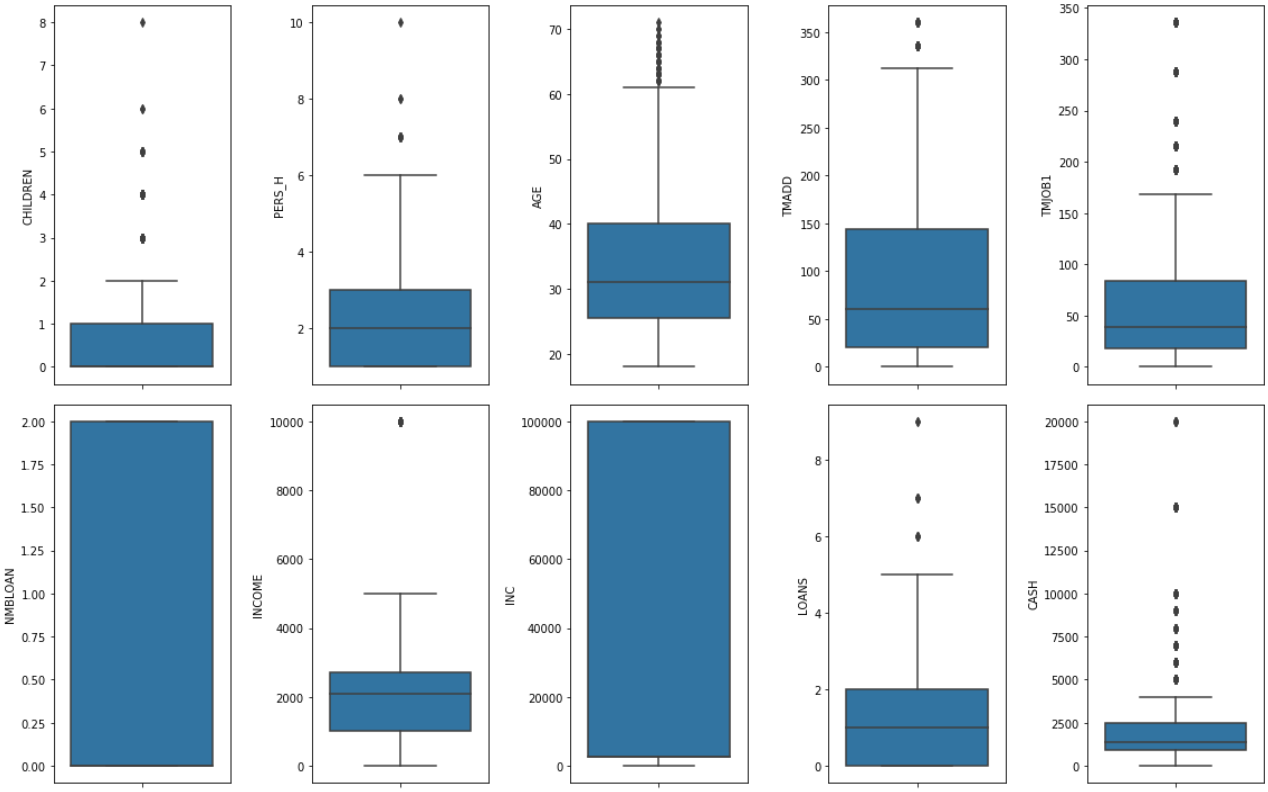

Figure 8 Plot after outlier processing

It is noticeable that other outliers are also present on it, but their removal from the analysis can lead to the loss of a large amount of useful data.

The frequency diagram of categorical variables (Fig. 9) shows that the variables PRODUCT, NAT, CAR, CARDS and TEL contain categories of low frequency.

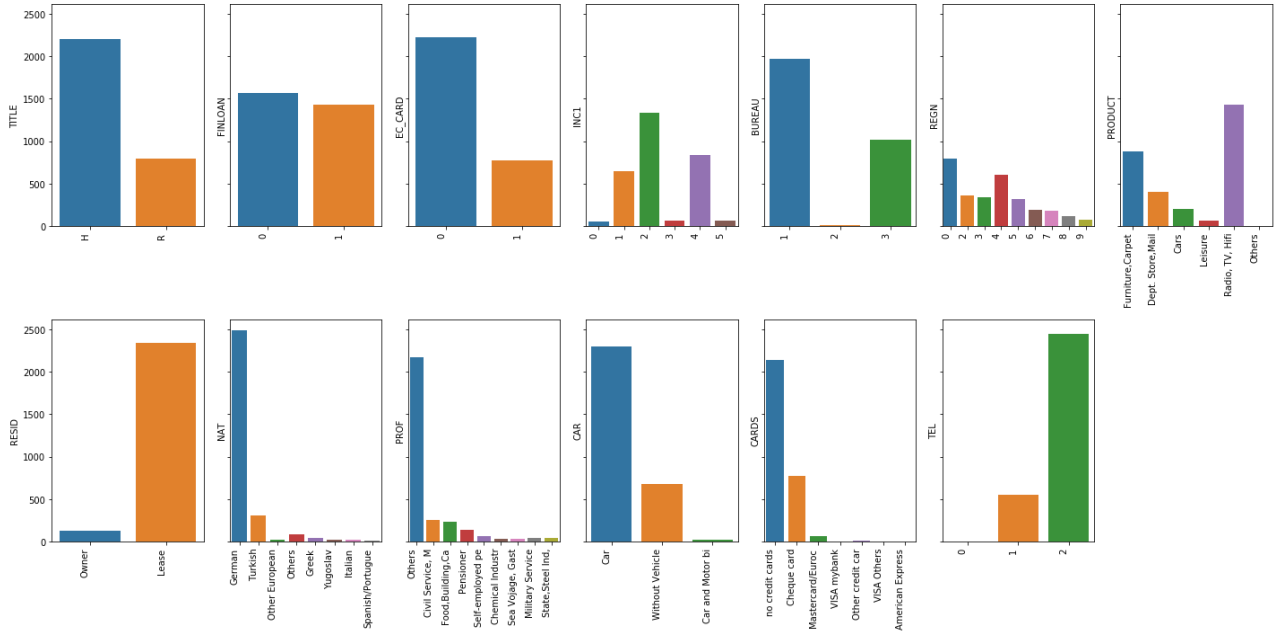

Figure 9 Raw data frequency diagram

Processing passes

The original sample contains 687 gaps. The gaps are localized in the variables TMADD, TMJOB, CASH, PRODUCT, RESID and PROF (Fig. 10, 11). The reasons for their appearance are most likely the lack of the necessary data at the collection stage. Since the variables contain up to 50% of the gaps, the strategy of filling them was applied. A separate category has been created in the RESID variable with 535 gaps. The existing Other category has been assigned to the missing values in the PRODUCT and PROF variables. The missing values of the numeric variables

TMADD, TMJOB, CASH are replaced with an average.

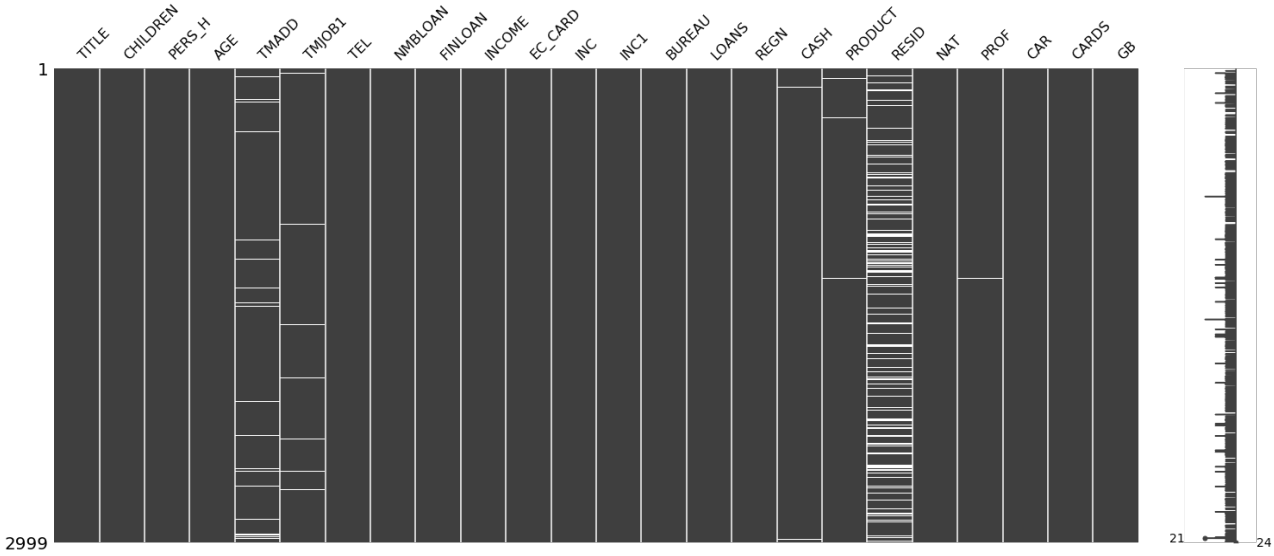

Figure 10 Missing data structure

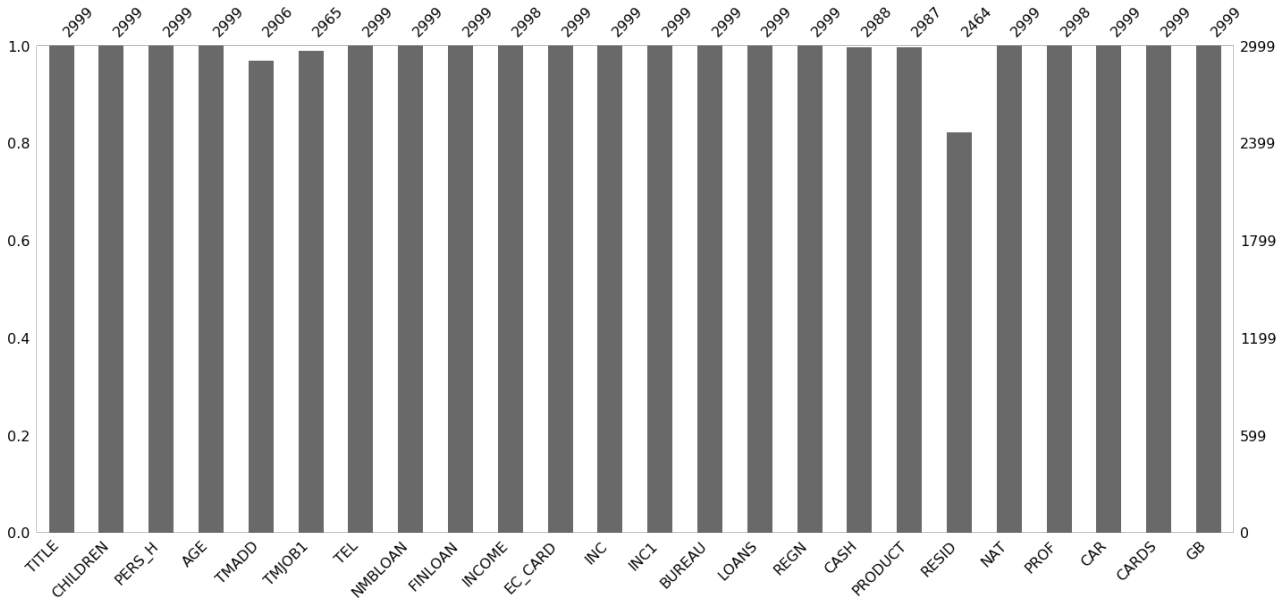

Figure 11 Number of gaps in data for variables

3.3 Selecting Variables and Test Accuracy

The important part to create the credit risk model is the variable reduction by using statistical methods SAS. Among the variety of statistical methods that are employed to analyze data set very popular regression methods. There are widely used in examining the relationship between target variable and other predictor variables. For class of regression methods, logistic regression is well suited for studying categorical variables. Logistics regression also helpful choosing the most important variables of the predictive model to make the model more accurate.

We have 25 variables. Our target variable is 'GB'(Good/Bad) which has 1 and 0. 1 for Good and 0 for Bad.

Category variables: car_, cards_, nat_, product_, prof_, rsid_, title_, GB.

Numeric Variables: CHILDREN, PERS_H, AGE, TMADD, TMJOB1, TEL, NMBLOAN, FINLOAN, INCOME, EC_CARD, INC, INC1, BUREAU, LOANS, REGN, CASH.

We have to use variable selection of logistic regression method PROC LOGISTIC (SAS) [7-9] to select the importance variable the effect to the our target variable the most, but before apply it to logistic regression model. We have to do the data partition first.

Before you do the variable selection step, you have to partition the data first. Splitting the dataset into training and validation by using the 70:30 ratio. First, I need to sort out the data using proc sort and splitting by using proc Surveyselect.

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	227.5907		<.0001	AGE
2	EC_CARD		1	2	85.9125		<.0001	EC_CARD
3	TEL		1	3	66.1721		<.0001	TEL
4	TMJOB1		1	4	33.8726		<.0001	TMJOB1
5	LOANS		1	5	18.1220		<.0001	LOANS
6	NMBLOAN		1	6	22.0734		<.0001	NMBLOAN
7	PERS_H		1	7	16.8270		<.0001	PERS_H
8	CHILDREN		1	8	35.5184		<.0001	CHILDREN
9	INC		1	9	2.2900		0.1302	INC
10	INC1		1	10	7.0426		0.0080	INC1

Figure 12 Result of proc logistic

Finally, as we can see, from 23 initial variables algorithm selected only 10 most important variables: AGE, EC_CARD, TEL, CHILDREN, TMJOB1, NMBLOAN, LOANS, INC, INC1, and PERS_H. Another 13 variables is not so important because they influence on target variable "GB" not significant.

You can see the probability (PR>ChiSq) value from the result of the model's table above that the model will selecting the variable depends on the PR>ChiSq value, so this value is often set the value between 0.005 or 0.01(Can call PR>ChiSq as p value). The p-value tested from the test, we confirmed that at least one of the regression coefficients in the model is not equal to zero.

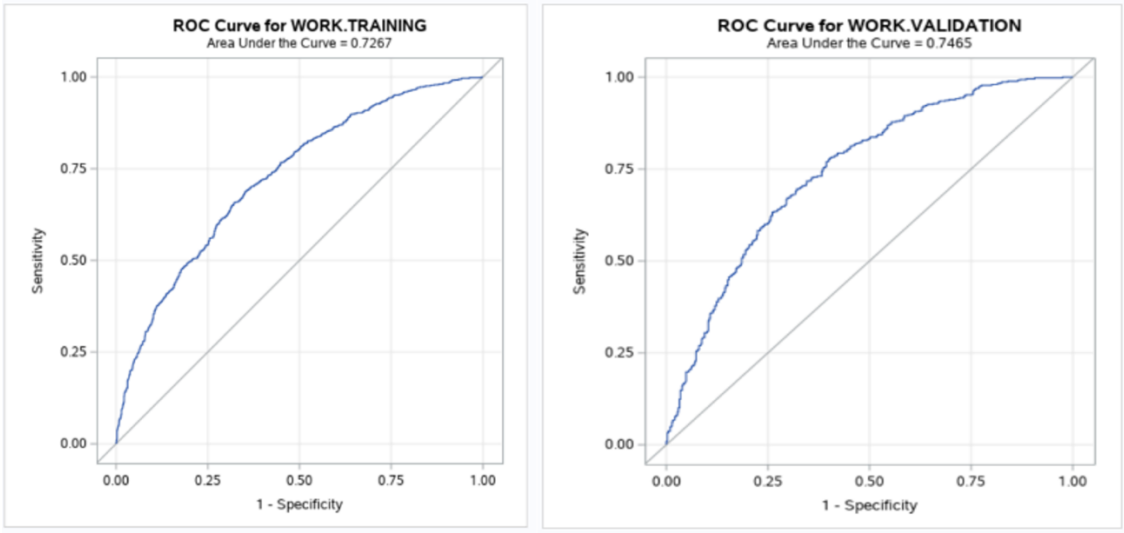

Figure 13 ROC Curve of how accurate of the dataset

After analysis initial data, using Logistic regression model, our predict model show good result – ROC curve for validation data more than 70%. We can see that the variable selection algorithm decided that the model would include CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD, INC, INC1 and LOANS variables(10 variables). In addition, our results are 72.67% accuracy for the TRAINING dataset and 74.65% accuracy for the VALIDATION dataset. In the next step we will use these 10 variables to do WOE and calculate credit risk score.

3.4 Data transformation

The range of values for numeric variables was divided into ten parts by proc rank and using WOE to scale the score. The table below will show the result of WOE an IV for each variable.

Table 4 Chidren table

CHILDREN	total_good	total_bad	woe	iv	good%	group_chil	score
2	707	877	-0.216144552	0.024630116	44.63384	0	7
6	386	290	0.285288221	0.018258094	57.10059	1	10
8	298	230	0.258345952	0.011708621	56.43939	2	9
9	106	99	0.067651018	0.000313347	51.70732	>=3	8

Table 5 AGE table

AGE	total_good	total_bad	woe	iv	good%	group_age	score
0	84	254	-1.107185694	0.125858085	24.85207	18-22	1
1	99	180	-0.598505227	0.032432155	35.48387	23-24	3
2	98	180	-0.608657598	0.033388883	35.2518	25-26	2
3	112	137	-0.20215028	0.00338829	44.97992	27-28	4
4	185	190	-0.027336473	9.36234E-05	49.33333	29-31	5
5	142	127	0.110971745	0.001105648	52.7881	32-34	6
6	190	128	0.394325582	0.016308916	59.74843	35-38	8
7	162	111	0.377397908	0.012838538	59.34066	39-42	7
8	206	114	0.591009495	0.03629114	64.375	43-50	9
9	219	75	1.07091539	0.102978041	74.4898	51-70	10

Table 6 Salary table

INC	total_good	total_bad	woe	iv	good%	group_inc	score
1	466	240	0.662878485	0.100002801	66.00567	0-2499	6
4	547	845	-0.435556051	0.086868218	39.29598	2500-9999	8
8	484	411	0.162823466	0.007910074	54.07821	>10000	10

Table 7 Salary + Ec_card from all the bank

INC1	total_good	total_bad	woe	iv	good%	group_inc1	score
0	38	20	0.64118566	0.007703921	65.51724	0	8
1	428	220	0.664827423	0.092308841	66.04938	1	9
4	502	826	-0.49866288	0.108110958	37.8012	2	5
6	45	19	0.861555285	0.014956243	70.3125	3	10
8	448	386	0.148287637	0.006115947	53.71703	4	6
9	36	25	0.363974888	0.002670435	59.01639	5	7

Table 8 Number of running loans

LOANS	total_good	total_bad	woe	iv	good%	group_loar	score
1	596	581	0.024821684	0.000242275	50.63721	0	8
5	536	507	0.054954932	0.00105215	51.39022	1	10
8	249	251	-0.008668269	1.25524E-05	49.8	2	6
9	116	157	-0.30332384	0.008328731	42.49084	>=3	4

Table 9 Number of loans that you have done in the past

NMBLOAN	total_good	total_bad	woe	iv	good%	group_nmb	score
3	994	1074	-0.078076294	0.004209857	48.06576	0	8
7	33	75	-0.821648778	0.023079787	30.55556	1	6
8	470	347	0.302739689	0.024827495	57.52754	>=2	10

Table 10 people in your household

PERS_H	total_good	total_bad	woe	iv	%good	group_pers	score
1	414	702	-0.528735656	0.101886425	37.09677	1	2
4	378	265	0.354496144	0.026716947	58.78694	2	10
6	328	243	0.299283939	0.016960936	57.44308	3	6
8	277	200	0.325031914	0.016689381	58.07128	4	8
9	100	86	0.150154664	0.001398486	53.76344	>5	4

Table 11 contract number

TEL	total_good	total_bad	woe	iv	%good	group_tel	score
0	204	349	-0.537620154	0.052157878	36.88969	1	8
5	1293	1147	0.119147036	0.011559196	52.9918	2	10

Table 12 Time at your current job

TMJOB1	total_good	total_bad	woe	iv	%good	group_tmj	score
0	80	152	-0.642522112	0.030946476	34.48276	0-8	1
1	122	203	-0.50985316	0.027633461	37.53846	9-17.	2
2	157	203	-0.2576284	0.00793979	43.61111	18-23	3
3	142	178	-0.226624719	0.005467906	44.375	24-32	4
4	112	99	0.122710795	0.0010602	53.08057	33-38	6
5	162	172	-0.060566367	0.000409237	48.50299	39-48	5
6	160	159	0.005601387	3.34406E-06	50.15674	49-71	7
7	181	145	0.221095063	0.0053026	55.52147	72-119	8
8	151	96	0.452263419	0.016596838	61.1336	120-190	9
9	230	89	0.948774713	0.089325845	72.10031	>191	10
						months	

Table 13 EC_Card (0/1)

EC_CARD	total_good	total_bad	woe	iv	good%	group_ec_c	score
3	988	1232	-0.221379672	0.036205045	44.5045	0	8
8	509	264	0.655830688	0.107256368	65.84735	1	10

From all the tables that you see from above, I will explain each column.

Column no.1 - number of group that got group by proc rank in SAS

Column no.2 - number of total good borrowers for each group

Column no.3 - number of total bad borrowers for each group

Column no.4 – WOE value for weight the score

Column no.5 – IV value to tell us how strong of predictive power

Column no.6 – percentage of good borrowers for each group

Column no.7- rang and group of value for each group

Column no.8 – score that we calculate follow the WOE value

After we get the scorecard I will show you the Best borrower's profile and the worse borrower's profile.

Table 14 the best and the worse borrower's profiles

	Age	income	inc 1	Tel num	loans	childrens	Tmjob (month)	Ec_ card	Num_loan	Person in house
Best score	51-70	>10000 $	3	>2	1	1	>191	1	>2	2
Worse score	18-22	0-2499 $	4	1	>3	0	0-8	0	1	1

3.5 Calculator (GUI)

After we weight all the score and have scorecard. Easy way to use a scorecard is create a calculator. I will create a calculator using python.

Figure 14 interface of the calculator and result

You can see the calculator from above that you can input the data of the 10 variables that we have selected from the logistic regression and press “Calculate score” the you will get the how good that profile will be in percentage, above 75% mean you can trust that profile.

4 FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING

In this work, the development of project management software modules for create an automation scorecard. The purpose of this section is to provide a business case of this development, as well as the definition and calculation of labor and money costs for its creation.

4.1 Assessment of the commercial potential and prospects of research from the perspective of resource efficiency and resource saving

4.1.1 Product consumer description

The developed software modules are intended for implementation and

internal use of the Bank to filter the new customer that want to loan money from the Bank and to improve the quality of work with clients through system automation that will decide the quality of the customer before giving them the loan and also solve the problem with machine learning for the customer that have the complex profile.

Separate commercial use of the developed modules, there can sale it to third parties, it is not expected, therefore further analysis is made from the standpoint of the convenience of the internal use and reduce the time spent by company employees orders.

4.1.2SWOT analysis

SWOT analysis is one of the most commonly used analysis methods in management and marketing. This method gives a clear idea of the current situation, and also helps to understand what actions need to be taken to maximize the project's capabilities and neutralize weaknesses and threats.

The purpose of using SWOT analysis for this development is determination of possible effectiveness and forecasting of directions future development of the developed solution.

The advantage of SWOT analysis is the development of connections various factors of the external and internal development environment.

The results of the SWOT analysis are presented in a summary Table 1, where the strengths and weaknesses of the development are indicated, possible directions for the future development of software modules are identified, and options for minimizing the impact of threats are considered.

Table 15

	Strengths of the project: 1.Automatic calculation 2. The use of modern technologies in the development process.	**Weaknesses of the project:** 1.Strictly defined structure of approval customer

	3.Use machine learning to solve the complex task. 4. Possibility of spot check of work.	2.Use a lot of data to make machine learning more accurate. 3. Binding to the Data just for each Bank.
Capabilities: 1.Use to create an expert system. 2. Apply this system to other Bank. 3. Formation of additional reports.	1) Creation of an expert system. 2. Implementation of system notification of management with proposed solutions. 3) Generation of additional reports for selective verification of the results.	1.Changing the implementation for a different system configuration. 2.Changing the method to collect the data 3.Changing Data to external Data to the system.
Threats: 1.System malfunctions 2.Do not have enough data to apply to machine learning.	1. Maintaining registers of accumulation of customer data. 2. Changing method to collect data	1. Changing the structure of data, reducing the critical elements (for the sake of flexibility). 2.Can also use some data of the other bank or old data.

4.2 Project initial

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.

4.2.1 The structure of work in the framework of scientific research

At this part, a complete list of necessary work is drawn up, the performer is appointed and the duration is set. The result work scheduling is a linear timeline for project implementation.

The list of stages of work and the distribution of performers is presented in Table 2.

Table 16 List of stages of work and distribution of performers

No.	Stages of work	Performers
1	Setting goals and objectives	Supervisor
2	Development and approval of technical specifications	Supervisor, Student
3	Selection and study of materials on the topic	Supervisor, Student
4	Development of a calendar plan	Supervisor, Student
5	Discussion of literature	Supervisor, Student
6	Analyzing the subject area	Student
7	Design	Supervisor, Student
8	Development	Student
9	Testing and Debugging	Student
10	Execution of the settlement and explanatory note	Student

Table 17 Stakeholders of the project

Project stakeholders	Stakeholder expectations
The Bank that wants to use automate scorecard	-Use scorecard instead of employees -Scorecard that accurate enough to use to decide newcomers.

Table 18 Purpose and results of the project

Purpose of project:	Create an automation scorecard
Expected results of the project:	Scorecard that can decide the reputation of the customer for loaning money from the bank.
Criteria for acceptance of the project result:	Fast and more accurate in calculate customer score.
Requirements for the project result:	-Dataset -SAS programming

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.

			Duration of the project															
		MONTH	FEBUARY				MARCH				APRIL				MAY			
ACTIVITY	PATICIPANTS	Tc, DAYS	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Setting goals and objectives																		
Development and approval of technical specifications																		
Selection and study of materials on the topic																		
Development of a calendar plan																		
Discussion of literature																		
Analyzing the subject area																		
Design																		
Development																		
Testing and Debugging																		
Execution of the settlement explanatory note																		
Graphic design																		

4.3 Project budget

For a project to develop software modules for project management costs are estimated for the following items:

- materials and purchased items
- salary
- social tax
- electricity costs (without lighting)
- depreciation charges
- other expenses

Since the work on the project was carried out without the involvement of third parties organizations and for its implementation did not require the rent of any property, and there was no need for business trips, expenses for there are no corresponding articles.

4.3.1 Calculation of material costs

In material costs, only the costs of stationery and printer cartridges are taken into account, since all the materials necessary for working on the project were at the disposal of the performers. Materials required to carry out this work, and the calculation material costs are presented in table 5.

Energy costs are calculated by the formula: ,

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

where P − power rates (5.8 rubles per 1 kWh for Tomsk); el

P − power of equipment, kW;

F_{eq} − equipment usage time, hours.

The nominal power of the personal computer is 0.3 kW.Assuming that the duration of the working day is 8 hours, and the work was performed 112 working days, we get that the total lead time the project is 896 hours. Since the work on the computer was carried out 7 hours a day out of 8. Then out of 896 hours spent by the performer

on the implementation project, 788.5 hours were spent at the computer. Electricity costs at operation of the equipment are summarized.

C = 5.8*0.3*788.5 = 1371.99 rubles

Table 19 Material costs

Name of materials	Price per unit, rub.	Qty	Amount, rub.
Printer paper, A4	240,00	1	240.00
Ballpoint pen	20,00	2	40.00
Cartridge	1500,00	1	1500.00
Notebook	50,00	2	100.00
Energy cost	5.8	236.55	1371.99
Total:			**3251.99**

4.3.2 Basic salary

This point includes the basic salary of participants directly involved in the implementation of the 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 following formula:

$$S_b = S_a \cdot T_w ,$$

where Sb – basic salary per participant;

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

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

The average daily salary is calculated by the formula:

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

Where 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; at holiday in 24 days, M = 11.2 months, 5 day per week;

F_V – valid annual fund of working time of scientific and technical staff.

The calculation of basic wages is performed on the basis of labor intensity completion of each stage and the amount of the monthly salary of the performer. Monthly salary of the scientific supervisor in the position Associate Professor and with a PhD in Technical Sciences, is 23264 rubles / month, the performer who is a student is 2200 rubles / month. Based on the fact that in a month on average 24.83 working days with a six-day working week, the average daily .

For accounting in its composition of bonuses, additional wages and regional allowances the district coefficient Kp = 1.3 is used.

Thus, for the transition from the tariff amount of earnings the contractor associated with participation in the project to the corresponding full earnings need to take into account the regional coefficient Kp = 1.3.

Table 20 The cost of wages

performer	**Salary, RUB /month**	**Average daily rate, rub/ workday**	**Time spent working days**	**Kp**	**Salary fund, rub.**
scientific adviser	23264.00	936.93	38	1.3	46284.34
student	2200.00	88.6	112	1.3	12900.16
Total:					**59184.50**

4.3.3 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:

$W_{add} = k_{extra} \cdot W_{base}$, (x) where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient; W_{base} – base salary, rubles.

Contributions to extra-budgetary funds include contributions to the pension fund, social and health insurance and account for 30.2% of the salaries of the project participants, the student scholarship is not taken into account.

$$W_{add} = 46284.34 * 0.302 = 13977.87$$

4.3.4Costs of special equipment

Depreciation allowances for the project under consideration include depreciation of the equipment used during the execution of the work. Depreciation deductions are calculated based on the time of use. Annual depreciation of NA is determined as the reciprocal of the amortization period of NA equipment, which is determined in accordance with the decree of the Government of the Russian Federation "On the classification of fixed assets included in depreciation groups." For a computer, let's take CA = 3 years, then NA = 0.33. For a printer, let's take CA = 2 years, then NA = 0.5.

Table 21 Depreciation expense

Equipment name	**Amortization rate. Equipment,NA**	**Price, rub.**	**operating time of computing equipment**	**effective annual working hours**	**Amortized deduction ,rub**
Personal Computer	0.33	45000.00	788.5	2384	4911.6
Laser printer	0.5	12000.00	15	2384	37.6
Total					**4949.2**

4.3.5Overhead 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:

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

This section estimates the costs of implementing projects that were not taken into account in previous articles, payment for communication services, copying, purchasing materials, etc. The amount of other expenses is 10% of the amount of all previous costs .

Cov = 0.1*(3251.99+13977.87+4949.2) = 8136.3

Thus, other overhead costs amounted to 8136.3 rubles.

4.3.6Formation 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

Name	Cost, rubles
1. Material costs and Energy cost	3251.99
2. Costs of special equipment	4949.2
3. Basic salary	59184.50
4. Additional salary	13977.87
5. overhead costs	8136.3
Total	**89499.86**

The total cost of the project was equal to 89499.86 rubles.

Conclusion

During the implementation of the financial management section, a comprehensive description and analysis of the financial and economic aspects of the work performed. A list of the work carried out, their performers and the duration of the work stages have been compiled, a line schedule has been drawn up. Also, the cost

estimate for the project was calculated, the cost of the project was calculated, the performance indicators of the project were determined and its effectiveness was assessed.

5 Social responsibility

Introduction

The developed project is to reduce the risk of the bank to make a decision for new borrowers using machine learning and SAS programming. The development of the program was carried out only 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.

Final qualification work on the development of project management software modules for create Automation scorecard The projected workplace is an room space in which the developer will work.

Room characteristics:

- working space width - 5 m, length - 6 m, height - 3.5 m.
- room area - 30 m2.
- room volume - 105 m3.
- an air conditioner is installed in the room, there is a natural ventilation - exhaust vent, door, window.
- artificial lighting is installed in the room, there is daylight.

5.1 Legal and organizational issues of occupational safety

Nowadays one of the main way 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 systemof 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.

According to the GOST 12.2.032-78 SSBT [1], every employee has the right:

- To have a workplace that meets Occupational safety requirements;
- To have a compulsory social insurance against accidents at manufacturingand occupational diseases;
- To receive reliable information from the employer, relevant government bodies and public organizations on conditions and Occupational safety at the workplace, about the existing risk of damage to health, as well as measures toprotect against harmful and (or) hazardous factors;
- To refuse carrying out work in case of danger to his life and health due to violation of Occupational safety requirements;
- Be provided with personal and collective protective equipment in compliance with Occupational safety requirements at the expense of the employer;
- For training in safe work methods and techniques at the expense of the employer;
- For personal participation or participation through their representatives in consideration of issues related to ensuring safe working conditions in his workplace, and in the investigation of the accident with him at work or occupational disease;
- For extraordinary medical examination in accordance with medical recommendations with preservation of his place of work (position) and secondary earnings during the passage of the specified medical examination;

- For warranties and compensation established in accordance with this Code, collective agreement, agreement, local regulatory an act, an employment contract, if he is engaged in work with harmful and (or) hazardous working conditions.

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.

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.

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

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 [2].

5.3 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 othermeasures.

Working conditions - a set of factors of the working environment and the laborprocess 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 thatcan 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 1 - Potential hazardous and harmful production factors

Factors (**GOST 12.0.003-2015**)	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Excessive levels of noise, vibration	+	+		**GOST 12.1.003-2014** Occupational safety standards system. Noise. General safety requirements
2. Insufficient illumination	+			**SanPiN 2.2.1/2.1.1.1278-03** Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings
3. Electromagnetic fields	+	+	+	**SanPiN 2.2.4.1329-03** Requirements for protection of personnel from the impact of impulse electromagnetic fields
4. Abnormally high voltage value in the circuit, the closure which may occur through the human body		+	+	**Sanitary rules GOST 12.1.038-82 SSBT.** Electrical safety. Maximum permissible levels of touch voltages and currents.

5.3.1 Excessive levels of noise, vibration

Noise and vibration worsen 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 result in weakened attention, deteriorated memory, decreased response, and increased number of errors in work.

Noise can be generated by operating equipment, air conditioning units, daylight illuminating devices, as well as spread from the outside.

When working on a PC, the noise level in the workplace should not exceed 50 dB [3].

5.3.2 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. 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 [4].

According to the SanPiN 2.2.2 / 2.4.1340-03 [4] standard, the illumination on the table surface in the area ofthe 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%.

5.3.3 Electromagnetic fields

In this case, the sources of increased intensity of the electromagnetic field are a personal computer. 8- 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 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 [5]:

- 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.

When working with a computer, the ionizing radiation source is a display. Under the influence of ionizing radiation in the body, there may be a violation of normal blood coagulability, an increase in the fragility of blood vessels, a decrease in immunity, etc. The dose of irradiation at a distance of 20 cm to the display is 50 μrem/hr. According to the norms [8], the design of the computer should provide the power of the exposure dose of x–rays at any point at a distance of 0,05 m from the screen no more than 100 μR/h. Fatigue of the organs of vision can be associated with both insufficient illumination and excessive illumination, as well as with the wrong direction of light.

5.3.4 Abnormally high voltage value in the circuit

The mechanical action of current on the body is the cause of electrical injuries. 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) [6].

5.4 Ecological 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 paper waste - 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 [8].

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 [8].

5.5 Safety in emergency

In the working environment of the PC operator, the following manufactured emergencies may occur [9]:

- 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 as a 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[12].

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[11].

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) [11].

Conclusion

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.

Conclusion

In this project have created the Credit risk model using SAS and python programming including the following step: prepare data, partition data, variable selection, test accuracy, calculate WOE and IV values and create the calculator interface (GUI). So proc logistic[7-9] have got 10 variables out of 23 variables and this project use that 10 variables to create calculator to support the decision making of the bank. The bank can use this calculator to control the risk of borrowers who want to loan money from the bank and also delete the bad borrowers and indicate good borrowers for the bank. This calculator also let you know which variable in your profile that you should take care of and the range of a good borrower should be.

Reference for social responsibility

1. GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements.
2. SanPiN 2.2.2 / 2.4.1340-03. Sanitary-epidemiological rules and standards "Hygienic requirements for PC and work organization".
3. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.
4. SanPiN 2.2.1 / 2.1.1.1278-03. Hygienic requirements for natural, artificial and combined lighting of residential and public buildings.
5. SanPiN 2.2.2 / 2.4.1340-03 "Hygienic requirements for personal computers and work organization "
6. GOST 12.1.038-82 Occupational safety standards system. Electrical safety
7. Federal Law "On the Fundamentals of Labor Protection in the Russian Federation" of 17.07.99 № 181 – FZ
8. GOST R ISO 1410-2010. Environmental management. Assessment of life Cycle. Principles and structure.
9. GOST R12.1.004-85 Occupational safety standards system. Fire safety
10. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements
11. GOST Industrial equipment. General safety requirements to working places
12. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements

Reference

1. Fabio Wendling Muniz de Andrade & Abraham Laredo Sicsú (2008) A Credit Risk Model for Consumer Loan Portfolios, Latin American Business Review, 8: 3, 75-91
2. Amos Taiwo Odeleye, Paper 3554-2019, Developing a Credit Risk Model Using SAS
3. Anshu B. Data Preprocessing Techniques for Data Mining // Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets. New Delhi, 2011. P. 6.
4. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, Data Preprocessing for Supervised Leaning, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME 1 NUMBER 1 2006 ISSN 1306-4428
5. Huang Shan, Gubin E. Data cleaning for data analysis // Youth and modern information technologies: Proceedings of the XVI Intern. scientific - practical conference of students, graduate students and young scientists. Tomsk, 2018 - S. 387-389.
6. Salvador García, Julián Luengo, Francisco Herrera, Data Preprocessing in Data Mining, Springer International Publishing Switzerland 2015, Switzerland.
7. Kallunki J[1].-P., Broussard J., Boehmer E. Using SAS in Financial Research (2002)(en)(166s)
8. Allison, P. (2012). Logistic Regression using SAS (Theory & Applications), 2nd edition. SAS Institute, Cary, NC
9. Baesens B., Rosch D., & Schenle H. (2017). Credit Risk Analytics. Wiley & SAS Institute
10. Piatetsky G. Knowledge Discovery Nuggets [Electronic resource] // CRISP-DM, still the top methodology for analytics, data mining, or data science projects. 2014. P. 1. URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (accessed: 25.05.2021).
11. Umair Shafique and Haseeb Qaiser, A Comparative Study of Data Mining

Process Models (KDD, CRISP-DM and SEMMA),Department of Information Technology, University of Gujrat, Gujrat, Pakistan

12. Herman Jair Gómez Palacios*, Robinson Andrés Jiménez Toledo, Giovanni Albeiro Hernández Pantoja, Álvaro Alexander Martínez Navarro, A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change,Department of System Engineering, Mariana University, 520002, Colombia
13. Ng A. Machine learning yearning. 5th ed. deeplearning.ai, 2018. 116 p.
14. Daniel Berrar,Cross-validation,Data Science Laboratory, Tokyo Institute of Technology 2-12-1-S3-70 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
15. Måns Magnusson, Michael Andersen, Johan Jonasson, Aki Vehtari,Bayesian leave-one-out cross-validation for large data,Proceedings of the 36th International Conference on Machine Learning, PMLR 97:4244-4253, 2019.
16. Ioannis Tsamardinos, Elissavet Greasidou & Giorgos Borboudakis , Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation,Machine Learning volume 107, pages1895–1922 (2018)
17. Bee Wah,Yapa Seng, HuatOngbNor, Huselina Mohamed Husaina,Using data mining to improve assessment of credit worthiness via credit scoring models,Expert Systems with Applications,Volume 38, Issue 10, 15 September 2011, Pages 13274-13283
18. Peng Ding, Fan Li, Causal Inference: A Missing Data Perspective, Statist. Sci. 33(2): 214-237 (May 2018). DOI: 10.1214/18-STS645
19. James M. RobinsAndrea RotnitzkyDaniel O. Scharfstein, Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models, Part of the The IMA Volumes in Mathematics and its Applications book series (IMA, volume 116)F[]
20. Inkhireeva T.A. Data mining classification techniques for credit scoring in banks // Математическое и программное обеспечение информационных, технических и экономических систем: материалы VI международной

молодежной научной конференции, Томск, 24-26 мая 2018 г. - Томск: ТГУ, 2018 - С. 362-365

Appendix A Program cod for create credit risk model and calculator (GUI)

Prepare data

```
data work.importcopy;
set work.import;
run;

/*Descriptive Statistic for Data Analysis*/
proc means data=work.importcopy N MIN MAX MEAN MEDIAN STD MAXDEC=2;
        TITLE 'Descriptive Statistics for numeric';
run;

proc freq data=work.importcopy;
        TITLE 'Descriptive Statistics for char';
        TABLES car cards nat product prof resid title/
NOCUM NOPERCENT;
run;

/*1.Missing data*/
/*1.1Create format to group missing and nonmissing*/
proc format;
        value $missfmt ''='Missing' other='Not Missing';
        value missfmt .='Missing' other='Not Missing';
run;


proc freq data=work.importcopy; /* initial data*/
format _CHAR_ $missfmt.;        /*apply format for the duration of this PROC*/
tables _CHAR_ /missing missprint nocum nopercent;
```

```
format _NUMERIC_ missfmt.;
tables _NUMERIC_ /missing missprint nocum nopercent;
run;


/*3.Outliers of data*/
proc univariate data=work.importcopy robustscale plot;
var income age;
run;

proc freq data=work.importcopy;
where income>=1000;
tables income;
 run;


DATA work.importcopy;
SET work.import;
IF income >= 10000 THEN DELETE;
RUN;
/*4. Duplicate case*/
proc sort data=work.importcopy
        nodupkey out=NotDuplicate;
by id;
run;

/*5.Multicollinearity in the original data*/
proc princomp data=work.importcopy
        outstat=work.importcopy_stat noprint;
run;
```

```
/*6. Digitalization of Data*/
data work.importcopy;
set work.import;

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

if cards="no credit cards"  then cards_=1;
if cards="American Express" then cards_=2;
if cards="Cheque card"      then cards_=3;
if cards="Mastercard/Euroc" then cards_=4;
if cards="Other credit car" then cards_=5;
if cards="VISA Others"      then cards_=6;
if cards="VISA mybank"      then cards_=7;

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

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="Radio, TV, Hifi"   then product_=5;
if product="Others"            then product_=6;

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="Pensioner"          then prof_=5;
if prof="Sea Vojage, Gast"   then prof_=6;
if prof="Self-employed pe"   then prof_=7;
if prof="State,Steel Ind,"   then prof_=8;
if prof="Others"             then prof_=9;

if resid="Lease" then rsid_=1;
if resid="Owner" then rsid_=2;

if title="H" then title_=1;
if title="R" then title_=2;

drop car cards nat product prof resid title; /*delete Char variables*/
run;

/*Standardize variables Z-score*/
proc standard data=work.importcopy mean=0 std=1 out=work.zimportcopy;
var income cash inc;
run;

proc means data=work.zimportcopy;
var income cash inc;
run;
```

```
/*Regression with standadized data*/
proc reg data=work.zimportcopy;
model cash = income inc; /*How the right effect to the left.*/
run;
quit;

/*Standadize estimates in Proc Reg.*/
proc reg data=work.importcopy;
model cash = income inc/stb;
quit;
```

Logistic regreesion

```
 /* Split data into two datasets : 70%- training 30%- validation */
Proc Surveyselect data=work.importcopy out=split seed= 1234 samprate=.7 outall;
Run;

Data training validation;
Set split;
if selected = 1 then output training;
else output validation;
Run;

/*Logistic Regression*/
ods graphics on;
proc logistic data=work.importcopy descending;
class car_ cards_ nat_ product_ prof_ rsid_ title_ GB;
model GB(event='1') = CHILDREN PERS_H AGE TMADD TMJOB1 TEL NMBLOAN FINLOAN INCOME EC_CARD INC INC1 BUREAU LOANS REGN CASH
```

```
/ selection=stepwise slstay=0.15 slentry=0.15 stb;
score data=training out= Logit_Training fitstat outroc=troc;
score data=validation out= Logit_Validation fitstat outroc=vroc;
Run;
ods graphics on;

/*An entry significance level of 0.15, specified in the slentry=0.15 option, means a*/
/*variable must have a p-value < 0.15 in order to enter the model regression.*/
/*An exit significance level of 0.15, specified in the slstay=0.15 option, means */
/*a variable must have a p-value > 0.15 in order to leave the model*/
```

Calculate WOE and IV

```
/*WOE and IV*/
/*rank infor in table in to rank*/
proc rank data=work.importcopy groups=10 out=work.rank;
var CHILDREN PERS_H AGE TMJOB1 TEL NMBLOAN EC_CARD INC INC1 LOANS;
run;

/*TRANSREG MORALS Algorithm Iteration History for Identity(GB)*/
proc transreg data=work.rank;
model identity(GB) = monotone(CHILDREN PERS_H AGE TMJOB1 TEL NMBLOAN EC_CARD INC INC1 LOANS);
run;

proc sql;
create table work.rank1 as
select CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD, INC, INC1, LOANS, GB,
sum(GB=1) as total_bad, sum(GB=0) as total_good
```

```
from  work.rank;
quit;

proc sql;
create table woe as
select CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD, INC,
INC1, LOANS,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD,
INC, INC1, LOANS, GB,
total_bad,total_good
from work.rank1)
group by CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD,
INC, INC1, LOANS;
quit;
/*WOE of children*/
proc sql;
create table woe_children as
select CHILDREN,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select CHILDREN, GB,total_bad,total_good
from work.rank1)
group by CHILDREN;
quit;

/*WOE of PERS_H*/
```

```
proc sql;
create table woe_pers_h as
select PERS_H,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select PERS_H, GB,total_bad,total_good
from work.rank1)
group by PERS_H;
quit;

/*WOE of AGE*/
proc sql;
create table woe_age as
select AGE,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select AGE, GB,total_bad,total_good
from work.rank1)
group by AGE;
quit;

/*WOE of TMJOB1*/
proc sql;
create table woe_TMJOB1 as
select TMJOB1,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
```

```
from (select TMJOB1, GB,total_bad,total_good
from work.rank1)
group by TMJOB1;
quit;

/*WOE of TEL*/
proc sql;
create table woe_TEL as
select TEL,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select TEL, GB,total_bad,total_good
from work.rank1)
group by TEL;
quit;

/*WOE of NMBLOAN*/
proc sql;
create table woe_NMBLOAN as
select NMBLOAN,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select NMBLOAN, GB,total_bad,total_good
from work.rank1)
group by NMBLOAN;
quit;

/*WOE of EC_CARD*/
```

```
proc sql;
create table woe_EC_CARD as
select EC_CARD,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select EC_CARD, GB,total_bad,total_good
from work.rank1)
group by EC_CARD;
quit;

/*WOE of INC*/
proc sql;
create table woe_INC as
select INC,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select INC, GB,total_bad,total_good
from work.rank1)
group by INC;
quit;

/*WOE of INC1*/
proc sql;
create table woe_INC1 as
select INC1,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
```

```
from (select INC1, GB,total_bad,total_good
from work.rank1)
group by INC1;
quit;

/*WOE of LOANS*/
proc sql;
create table woe_LOANS as
select LOANS,
sum(GB=0) as total_good, sum(GB=1) as total_bad,
log((sum(GB=0)/sum(GB=1))/(mean(total_good)/mean(total_bad))) as woe,
(sum(GB=0)/mean(total_good)-sum(GB=1)/mean(total_bad)) * calculated woe as iv
from (select LOANS, GB,total_bad,total_good
from work.rank1)
group by LOANS;
quit;
```

Create calculator (GUI)

```
from tkinter import *
from tkinter import messagebox
fields = ('Age (18-70)', 'Salary', 'Salary+ec_card', 'Telephone', 'Num of running loans',
'Num of Children'
        , 'Time at Job (month)', 'EC_card holders(0/1)', 'Num Mybank Loans', 'Num in
Household')
def calculate_credit_score(entries):
   sum=0
   # age:
   age = float(entries['Age (18-70)'].get())
   print("age", age)
   if(age>=18 and age<=22):
      sum+=1
```

```
elif(age>=23 and age<=24):
    sum+=2
elif (age >= 25 and age <= 26):
    sum += 3
elif (age >= 27 and age <= 28):
    sum += 4
elif (age >= 29 and age <= 31):
    sum += 5
elif (age >= 32 and age <= 34):
    sum += 6
elif (age >= 35 and age <= 38):
    sum += 7
elif (age >= 39 and age <= 42):
    sum += 8
elif (age >= 43 and age <= 50):
    sum += 9
elif (age >= 51 and age <= 70):
    sum += 10
print("sum age:"+str(sum))
# salary:
salary = float(entries['Salary'].get())
if (salary >= 0 and salary <= 2499):
    sum += 6
elif (salary >= 2500 and salary <= 9999):
    sum += 8
elif (salary >= 10000):
    sum += 10
print("sum salary:" + str(sum))
# salary+ec_card:
salary_ec_card = float(entries['Salary+ec_card'].get())
```

```
if (salary_ec_card == 0):
    sum += 8
elif (salary_ec_card ==1):
    sum += 9
elif (salary_ec_card ==2):
    sum += 5
elif (salary_ec_card ==3):
    sum += 10
elif (salary_ec_card ==4):
    sum += 6
elif (salary_ec_card ==5):
    sum += 7
print("sum inc1:" + str(sum))
# Telephone:
tel = float(entries['Telephone'].get())
if (tel == 1):
    sum += 8
elif (tel >1):
    sum += 10
print("sum tel:" + str(sum))
#Num of running loans
loans = float(entries['Num of running loans'].get())
if (loans==0):
    sum += 8
elif (loans ==1):
    sum += 10
elif (loans == 2):
    sum += 6
elif (loans >= 3):
    sum += 4
```

```
print("sum loans:" + str(sum))
#Num of Children
nmChild = float(entries['Num of Children'].get())
if (nmChild == 0):
    sum += 7
elif (nmChild == 1):
    sum += 10
elif (nmChild == 2):
    sum += 9
elif (nmChild >= 3):
    sum += 8
print("sum nmChild:" + str(sum))
#Time at Job
tmbjob = float(entries['Time at Job (month)'].get())
if (tmbjob <= 8):
    sum += 1
elif (tmbjob >= 9 and tmbjob <= 17):
    sum += 2
elif (tmbjob >= 18 and tmbjob <= 23):
    sum += 3
elif (tmbjob >= 24 and tmbjob <= 32):
    sum += 4
elif (tmbjob >= 33 and tmbjob <= 38):
    sum += 5
elif (tmbjob >= 39 and tmbjob <= 48):
    sum += 6
elif (tmbjob >= 49 and tmbjob <= 71):
    sum += 7
elif (tmbjob >= 72 and tmbjob <= 119):
    sum += 8
```

```
elif (tmbjob >= 120 and tmbjob <= 190):
    sum += 9
elif (tmbjob > 190 ):
    sum += 10
print("sum tmjob:" + str(sum))
#EC_card holders
ec_card = float(entries['EC_card holders(0/1)'].get())
if (ec_card == 0 ):
    sum += 8
elif (ec_card == 1):
    sum += 10
print("sum EC_card holders:" + str(sum))
#Num Mybank Loan
nmbloan = float(entries['Num Mybank Loans'].get())
if (nmbloan==0):
    sum += 8
elif (nmbloan ==1):
    sum += 6
elif (nmbloan >= 2):
    sum += 10
print("sum nmbloan:" + str(sum))

#Num in Household
pers_h = float(entries['Num in Household'].get())
if (pers_h==1):
    sum += 2
elif (pers_h ==2):
    sum += 10
elif (pers_h == 3):
    sum += 6
```

```
    elif (pers_h ==4):
        sum += 8
    elif (pers_h >= 5):
        sum += 4
    print("sum pers_h:" + str(sum))

    print("Percentage of being a good customer:" +str(sum)+ "%")
    messagebox.showinfo("Percentage of being a good customer", "Percentage of being
a good customer : " +str(sum)+"%")

def makeform(root, fields):
  entries = {}
  for field in fields:
    row = Frame(root)
    lab = Label(row, width=22, text=field+": ", anchor='w')
    ent = Entry(row)
    ent.insert(0,"0")
    row.pack(side = TOP, fill = X, padx = 5 , pady = 5)
    lab.pack(side = LEFT)
    ent.pack(side = RIGHT, expand = YES, fill = X)
    entries[field] = ent
  return entries
if __name__ == '__main__':
  root = Tk()
  root.title("Scorecard")
  ents = makeform(root, fields)
  root.bind('<Return>', (lambda event, e = ents: fetch(e)))
  b2 = Button(root, text='Calculate score',
  command=(lambda e = ents: calculate_credit_score(e)))
  b2.pack(side = LEFT, padx = 5, pady = 5)
```

```
b3 = Button(root, text = 'Quit', command = root.quit)
b3.pack(side = LEFT, padx = 5, pady = 5)
root.mainloop()
```