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)

School: School of Engineering of Information Technology and Robotics
Field of training (Specialty): 09.04.04. Software engineering
Division: Big Data Solution

MASTER'S GRADUATION THESIS

Topic of research work
Data Mining Classification Techniques for Credit Scoring in Banks

UDC: 004.62:336.774.3

Student

Group	Full name	Signature	Date
8PM9I	Weijia Zhang		

Scientific supervisors

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	Gubin E. I	PhD		

ADVISERS:

Section “Financial Management, Resource Efficiency and Resource Saving”

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	Goncharova N. A	PhD		

Section “Social Responsibility”

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	Antonevich O. A	PhD		

ADMITTED TO DEFENSE:

Director of the programme	Full name	Academic degree, academic rank	Signature	Date
Data Mining	Sidorenko T. V	PhD		

Tomsk – 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)

School: School of Engineering of Information Technology and Robotics
Field of training (Specialty): 09.04.04. Software engineering
Division: Big Data Solution

APPROVED BY:
Director of the programme
___________Sidorenko T. V.
«____»_____________2021

ASSIGNMENT
for the Graduation Thesis completion

In the form:

Master's Dissertation

For a student:

Group	Full name
8PM9I	Weijia Zhang

Topic of research work:

Data Mining Classification Techniques for Credit Scoring in Banks	
Approved by the order of the Director of School of Information Tech & Robotics (date, number):	

Deadline for completion of Master's Graduation Thesis:	

TERMS OF REFERENCE:

Initial date for research work: *(Cleaning and mining of bank user data; establishment, mining, and accuracy verification of machine learning models; establishment of bank user credit score cards and user score grouping; operation characteristics of objects or products in terms of operational safety, environmental impact, and energy costs Special requirements; economic analysis, etc.)*	According to the bank's user history, the bank user credit score card is established, and the user is quantitatively analyzed according to various information of the user, and then the user is classified according to the user score according to the score interval.

List of the issues to be investigated, designed and developed *(Analytical review of literary sources with the purpose to study global scientific and technological achievements in the target field, formulation of the research purpose, design, construction, determination of the procedure for research, design, and construction, discussion of the research work results, formulation of additional sections to be developed; conclusions).*	1. The establishment of user credit score card. 2. A machine learning model suitable for user credit score cards. 3. The accuracy of the credit score card is verified, and customers are classified according to the credit score.

Advisors to the sections of the Master's Graduation Thesis
(With indication of sections)

Section	Advisor
1. Literature review	Gubin E. I
2. Practical part	Gubin E. I
3. Financial management	Goncharova N. A
4. Social Responsibility	Antonevich O. A

Date of issuance of the assignment for Master's Graduation Thesis completion according to the schedule	

Assignment issued by a scientific supervisors/advisor:

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	Gubin E. I	PhD		

Assignment accepted for execution by a student:

Group	Full name	Signature	Date
8PM9I	Weijia Zhang		

Tomsk – 2021

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

To the student:

Group	Full name
8PM9I	Weijia Zhang

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

Input data to the section «Financial management, resource efficiency and resource saving»:	
1. *Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human*	– Salary costs – 324906 – STR budget – 189885.7
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs –5,8 rub per 1 kW
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Labor tax –27,1 %; – Overhead costs –30%;
he list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	– comparative analysis with other researches in this field;
2. *Development of charter for scientific-research project*	– SWOT-analysis;
3. *Scheduling of STR management process: structure and timeline, budget, risk management*	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. *Resource efficiency*	– integral indicator of resource efficiency for the developed project.
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*	

ate of issue of the task for the section according to the schedule	22.02.2021

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	Goncharova. N. A	PhD		22.02.2021

The task was accepted by the student:

Group	Full name	Signature	Date
8PM9I	Weijia Zhang		22.02.2021

Tomsk – 2021

Task for section
«Social responsibility»

To student:

Group	Full name
8PM9I	Weijia Zhang

School	Information Techno & Robotics	**Department**	Information Technology
Degree	Master programmer	**Specialization**	09.04.04 Software Engineering

Title of graduation thesis:

Data Mining Classification Techniques for Credit Scoring in Banks	
Initial data for section «Social Responsibility»:	
1. Information about object of investigation (matter, material, device, algorithm, procedure, workplace) and areaof its application	– Build user credit score card based on user information – Use scoring cards to measure the default probabilities of users in different segments and set user loan credit scoring limits – This project was completed in the TPU dormitory at ycova 15b.
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	– Insufficient illumination of workplace – Excessive noise – Increased / decreased air humidity in the workplace; – Abnormally high voltage value in the circuit, the closure which may occur through the human body – Visual fatigue caused by using the computer for a long time – Neck soreness caused by prolonged sitting still
3. Ecological safety:	– Any harmful substances, that can emit into hydrosphere, atmosphere and lithosphere in case of unproper disposal or recycling of hazardous computer components
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		

Group	Full name	Signature	date
8PM9I	Weijia Zhang		

Expected learning outcomes in the direction

09.04.04 «Software Engineering»

Learning outcome code	Learning outcome (graduate must be ready)
General in the field of training 09.04.04 « Software Engineering »	
P1	Conduct scientific research related to the objects of professional activity
P2	Develop new and improve existing methods and algorithms for data processing in information and computing systems
P3	Prepare reports on the research work carried out and publish scientific results
P4	Design parallel processing systems and high-performance systems
P5	Implement software implementation of information and computing systems, including distributed
P6	Implement software implementation of systems with parallel data processing and high-performance systems
P7	Organize industrial testing of the created software
Big Data «Technology Profile» «Big data solutions»	
P8	Explore and analyze big data, create models of it, and interpret data structures in such models
P9	Understand the principles of creating, storing, managing, transferring and analyzing big data using the latest technologies, tools and data processing systems in high-performance networks
P10	Apply distributed database management system theory to traditional distributed relational database systems, cloud databases, large-scale machine learning systems, and data warehouses

Tomsk – 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)

School: School of Engineering of Information Technology & Robotics
Field of training (specialty): 09.04.04 «Software engineering»
Level of education: Master Degree Program
Division: Big data Solutions
Period of completion 2019/2020 and 2020/2021 academic years
Form of presenting the work:

Master's Thesis

SCHEDULED ASSESSMENT CALENDAR
for the Master's Graduation Thesis completion

Deadline for completion of Master's Graduation Thesis:	

Assessment date	**Title of section (module) / type of work (research)**	**Maximum score for the section (module)**
27.01.2021	1. Preparation of technical specifications and selection of research areas	
24.02.2021	2. Development of a common research methodology	
23.03.2021	3. Selection and study of materials on the topic	
13.04.2021	4. Obtaining necessary data and verification of the obtained results	
27.04.2021	5. Processing received data	
18.05.2021	6. Registration of the work performed	
29.05.2021	7. Preparation for defending a dissertation	

COMPILED BY:
Scientific supervisors:

Position	**Full name**	**Academic degree, academic rank**	**Signature**	**Date**
Associate professor	Gubin E. I	Ph.D		

Adviser

Position	**Full name**	**Academic degree, academic rank**	**Signature**	**Date**
Associate professor	Gubin E. I	Ph.D		

AGREED BY:
Director of the programme

Position	**Full name**	**Academic degree, academic rank**	**Signature**	**Date**
Associate Professor	Sidorenko T. V.	Ph.D		

Abstract

The work contains an explanatory note on 78 sheets, contains 18 figures, 18 tables, 1 application.

Key words: credit score card, weight of evidence, information value, logistic regression, customer’s segmentation.

Credit scoring technology is an applied statistical model whose function is to score loan applicants (credit card applicants) for risk assessment. The credit scoring card model is a mature forecasting method, especially in the areas of credit risk assessment and financial risk control. The credit score card can evaluate the customer's credit based on the information provided by the customer, the customer's historical data, and the data of the third-party platform. The establishment of the credit score card is based on the statistical analysis results of a large amount of data, which has high accuracy and reliability. This article uses the bank's customer data to establish a bank credit score card with high accuracy, and provides classification indicators to provide a good basis for the bank's customer segmentation.

Table of Contents

1. Project introduction:

Credit scoring technology is an applied statistical model, and its function is to make risk assessment scores for loan applicants (credit card applicants). The credit scoring card model is a mature forecasting method, especially in the fields of credit risk assessment and financial risk control [5]. The credit score card can evaluate the customer's credit based on the information provided by the customer, the customer's historical data, and the data of the third-party platform. The establishment of the credit score card is based on the statistical analysis results of a large amount of data, which has high accuracy and reliability.

The higher the user's credit score, the lower the user's default probability, and the bank's lending business is more secure, but this does not mean that the bank's lending business can get the most benefit. Between the best interests and business security, banks have to make a trade-off and carefully formulate lending standards [2].

This project mainly uses Python for analysis and modeling. Python has powerful data analysis and drawing capabilities. Using Python is also conducive to GUI development.

2. Dataset introduction:

In this project, the data comes from Kaggle (Give Me Some Credit). The dataset contains 150,000 credit information about customers applying for loans, including 11 variables [3]. The general data situation is shown in the following table.

	Variable Name	Description	Type
0	SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
1	RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
2	age	Age of borrower in years	integer
3	NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
4	DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
5	MonthlyIncome	Monthly income	real
6	NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
7	NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
8	NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
9	NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
10	NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

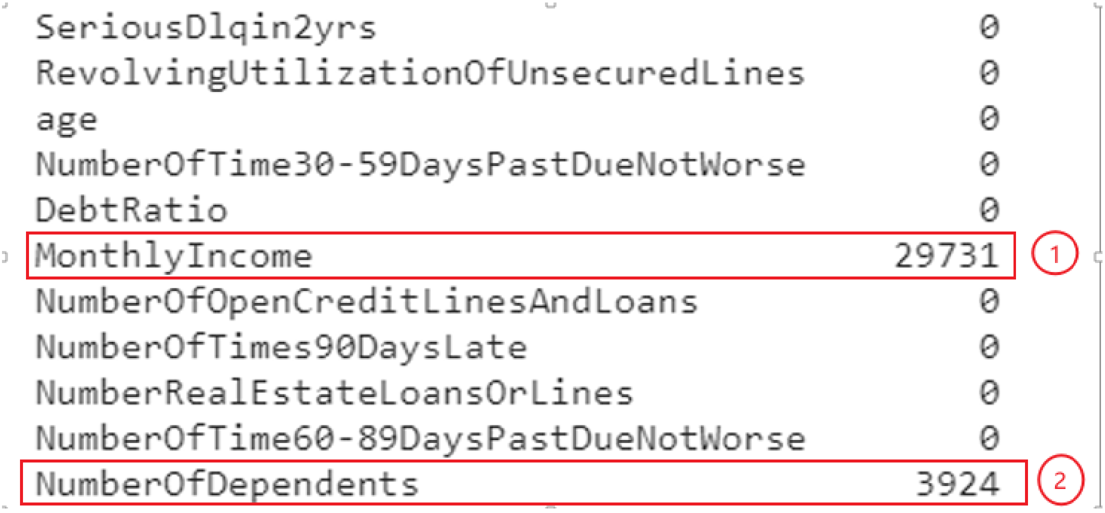
Table 1. Bank customer's dataset basic info

3.Prepare dataset

Before data analysis, the data set needs to be processed. The data set is generally repeated rows, noise values, noise labels, etc., which need to be corrected step by step for the problems of the data set [6]. If the machine learning model is used for analysis and prediction, it is necessary to divide the training set and Test set.

3.1Drop duplicates and handle NA value

Clean dataset, drop duplicates, drop useless columns or fill in columns. If the numbers of null rows are not so big, those rows don’t have a big influence for the whole dataset, can delete those rows directly. If the numbers of null rows are so big, those rows have a big influence for the whole dataset, for the accuracy of future analysis and the building of a better machine learning model, can fill the null rows using different ways. In this project using random freest to fill the null value.



```
SeriousDlqin2yrs                            0
RevolvingUtilizationOfUnsecuredLines        0
age                                         0
NumberOfTime30-59DaysPastDueNotWorse        0
DebtRatio                                   0
MonthlyIncome                           29731
NumberOfOpenCreditLinesAndLoans             0
NumberOfTimes90DaysLate                     0
NumberRealEstateLoansOrLines                0
NumberOfTime60-89DaysPastDueNotWorse        0
NumberOfDependents                       3924
```

Figure 1. Bank customer’s dataset na_value info

In this project, the dataset has 150 thousand data, NumberOfDependents has 3924 null values, it influences a little for the whole dataset. MonthlyIncome has almost 30 thousand data, it influences a lot for the whole dataset. In figure 1, for 2, just delete those null values rows, for 1, use random forest to fill the value.

Core code:

```
mData = data.iloc[:,[5,0,1,2,3,4,6,7,8,9]]
train_known = mData[mData.MonthlyIncome.notnull()].to_numpy()
train_unknown = mData[mData.MonthlyIncome.isnull()].to_numpy()
#Batch predict data and process
train_X = train_known[:,1:]
train_y = train_known[:,0]
rfr = RandomForestRegressor(random_state=0,n_estimators=200,max_depth=3,n_jobs=-1)
rfr.fit(train_X,train_y)
predicted_y = rfr.predict(train_unknown[:,1:]).round(0)
data.loc[data.MonthlyIncome.isnull(),'MonthlyIncome'] = predicted_y

data = data.dropna()
data = data.drop_duplicates()
data.isnull().sum()
```

3.2 Handle outliers in dataset

Outliers are objects that deviate from typical data. Strong outliers are considered anomalies, which are expected to be detected and analyzed further. They can represent significant information and need to be detected critically in many applications such as earth science, fraud detection, medical diagnosis, data cleaning, biological sequences, abnormal events from images and videos, and traffic movement patterns. They can also affect statistical analyses that are based on significance tests. Weak outliers are considered noise, which may harm data analysis such as clustering. In any case, regardless of strong or weak, outliers need to be detected [12].

Make a box plot to observe whether there are outliers in the data and handle those outliers. Outliers generally refer to values that deviate greatly from the data. For example, in statistics, outliers are defined as values less than Q1-1.5IQR or greater than Q3+1.5IQR. We observe the abnormal value of each variable by drawing a box plot and deal with it.

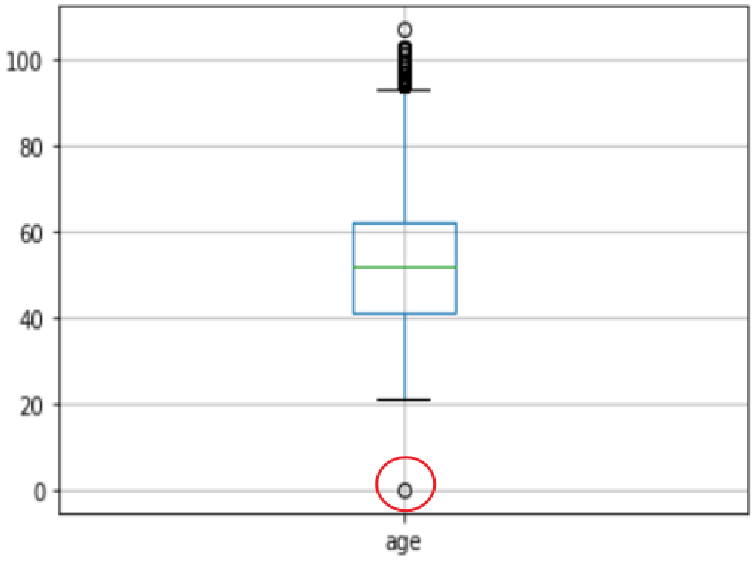


Figure 2. Age box plot outliers

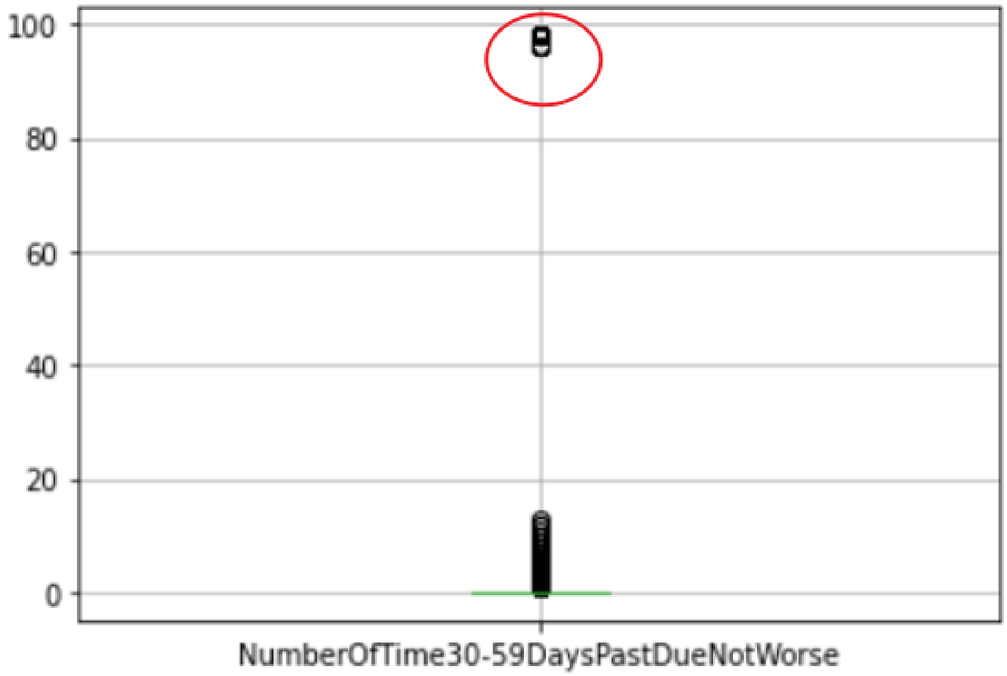


Figure 3. NumberRealEstateLoansOrLines box plot outliers

```
    Core code:

    data[['age']].boxplot()

    #delete age is 0

    data = data[data['age']>0]

    data[['NumberOfTime30-
59DaysPastDueNotWorse']].boxplot()

    #delete more than 20

    data = data[data['NumberOfTime30-
59DaysPastDueNotWorse']<20]
```

For example, in this project, as the figure 2 and figure 3 shown, the outliers of NumberOfTime30-59DaysPastDueNotWorse are those more than 80, the outliers of age are those less than 20. Then analyze and process other data columns according to this method.

3.3 Split the data

Data segmentation is to build a better machine learning model. By training on a subset of data, and testing on a different subset of data that the learning algorithm has never seen, ensure that the machine learning model is actually finding real patterns in the data and not just memorizing it.

We usually split the data around 20%-80% between testing and training stages, which performed on the figure 4. In this work, the dataset is so big, to prevent the data model from overfitting, we reduce the percentage of the training set and split a dataset into a training data (70%) and test data (30%).

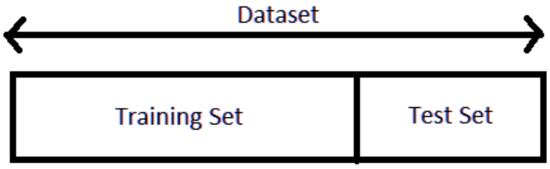


Figure 4. Train Set and Test Set split

```
    Core code:

    Y = data.iloc[:,0]

    X = data.iloc[:,1:]

    # Y=data['SeriousDlqin2yrs']

    # X=data.ix[:,1:]

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

    trainDf = pd.concat([Y_train, X_train], axis=1)

    testDf = pd.concat([Y_test, X_test], axis=1)

    clasTest =
testDf.groupby('SeriousDlqin2yrs')['SeriousDlqin2yrs'
].count()
```

4. Box handling

It is also called discretization of continuous variables. In the development of credit scoring cards, there are generally equidistant, equal frequency, and chi-square binning methods. First select the optimal segmentation for continuous variables, and then consider equidistant segmentation for continuous variables when the distribution of continuous variables does not meet the requirements of optimal segmentation.

The characteristics and continuity of the variable determine the type of binning of the variable. The binning is of great significance to future WOE calculations.

In this project, continuous variables can be optimized segmentation, and discontinuous variables can be manually binned. In this project, RevolvingUtilizationOfUnsecuredLines, age, DebtRatio and MonthlyIncome are optimized segmentation, and the remaining variables are manually binned.

For the features that cannot be reasonably split by the above binning method, manual binning without supervised binning is used.

5.WOE

WOE stands for Weight of Evidence. WOE is an encoding form of original independent variables. To perform WOE encoding on a variable, you need to group this variable first (also called discretization and binning). After grouping, for the i-th group, the WOE calculation formula is as follows:

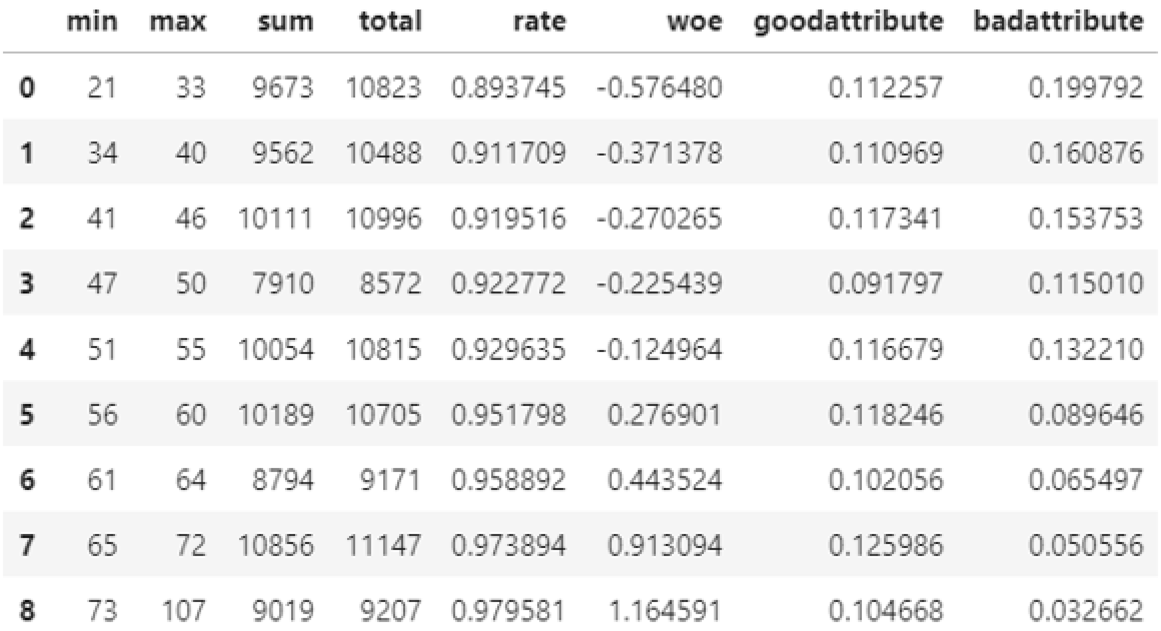
$$WOE_i = ln\left(\frac{p_{yi}}{p_{ni}}\right) = ln\left(\frac{y_i/y_T}{n_i/n_T}\right) = ln\left(\frac{y_i/n_i}{y_T/n_T}\right)$$

The above formula indicates that WOE is actually the difference between "the proportion of responding customers in the current grouping of all responding customers" and "the proportion of non-responding customers in the current grouping of customers who have not responded".

In this project, woe=ln(goodattribute/badattribute). The goodattribute calculation method is the number of good customers in each box/the total number of good customers in the data set; the badattribute calculation method is the number of bad customers in each box/the total number of bad customers in the data set.

Core code:

```
    def mono_bin(Y, X, n=10):
        r = 0
        good=Y.sum()
        bad=Y.count()-good
        while np.abs(r) < 1:
            d1 = pd.DataFrame({"X": X, "Y": Y,
"Bucket": pd.qcut(X, n)})
            d2 = d1.groupby('Bucket', as_index = True)
            r, p = stats.spearmanr(d2.mean().X,
d2.mean().Y)
            n = n - 1
        d3 = pd.DataFrame(d2.X.min(), columns =
['min'])
        d3['min']=d2.min().X
        d3['max'] = d2.max().X
        d3['sum'] = d2.sum().Y
        d3['total'] = d2.count().Y
        d3['rate'] = d2.mean().Y
        d3['woe']=np.log((d3['rate']/good)/((1-d3['rate'])/bad))
        d3['goodattribute']=d3['sum']/good
        d3['badattribute']=(d3['total']-d3['sum'])/bad
        iv=((d3['goodattribute']-d3['badattribute'])*d3['woe']).sum()
    #     print('d3 type')
    #     print(type(d3))
    #     print(d3.info())
    #     d3.loc['min'].sort_values(by = 'XXX')
    #     d3 dataframe sort_index how to use by and how to sort d3
        d4 = (d3.sort_values(by = 'min')).reset_index(drop=True)
    #     d4 = (d3.sort_index()).reset_index(drop=True)
        print("=" * 60)
        print(d4)
        woe=list(d4['woe'].round(3))
        cut=[]
        cut.append(float('-inf'))
        print(woe)
        for i in range(1,n+1):
            qua=X.quantile(i/(n+1))
            cut.append(round(qua,4))
        cut.append(float('inf'))
        return d4,iv,cut,woe
```



	min	max	sum	total	rate	woe	goodattribute	badattribute
0	21	33	9673	10823	0.893745	-0.576480	0.112257	0.199792
1	34	40	9562	10488	0.911709	-0.371378	0.110969	0.160876
2	41	46	10111	10996	0.919516	-0.270265	0.117341	0.153753
3	47	50	7910	8572	0.922772	-0.225439	0.091797	0.115010
4	51	55	10054	10815	0.929635	-0.124964	0.116679	0.132210
5	56	60	10189	10705	0.951798	0.276901	0.118246	0.089646
6	61	64	8794	9171	0.958892	0.443524	0.102056	0.065497
7	65	72	10856	11147	0.973894	0.913094	0.125986	0.050556
8	73	107	9019	9207	0.979581	1.164591	0.104668	0.032662

Figure 5. Age woe value

Figure 5 is the information after the age variable is subjected to the WOE binning operation. It can be seen from Figure 5 that the age variable is divided into 9 groups, and each group corresponds to a WOE value. As the WOE value increases, the proportion of bad customers decreases, which proves the accuracy of the WOE value.

6.Correlation analysis of variables

Through the heatmap, you can check the linear relationship between different variables and whether there is serious collinearity.

For the establishment of machine learning models, variables with strong collinearity should be proposed, and only one of the variables with strong collinearity should be retained. This can greatly improve the accuracy and interpretability of machine learning models.

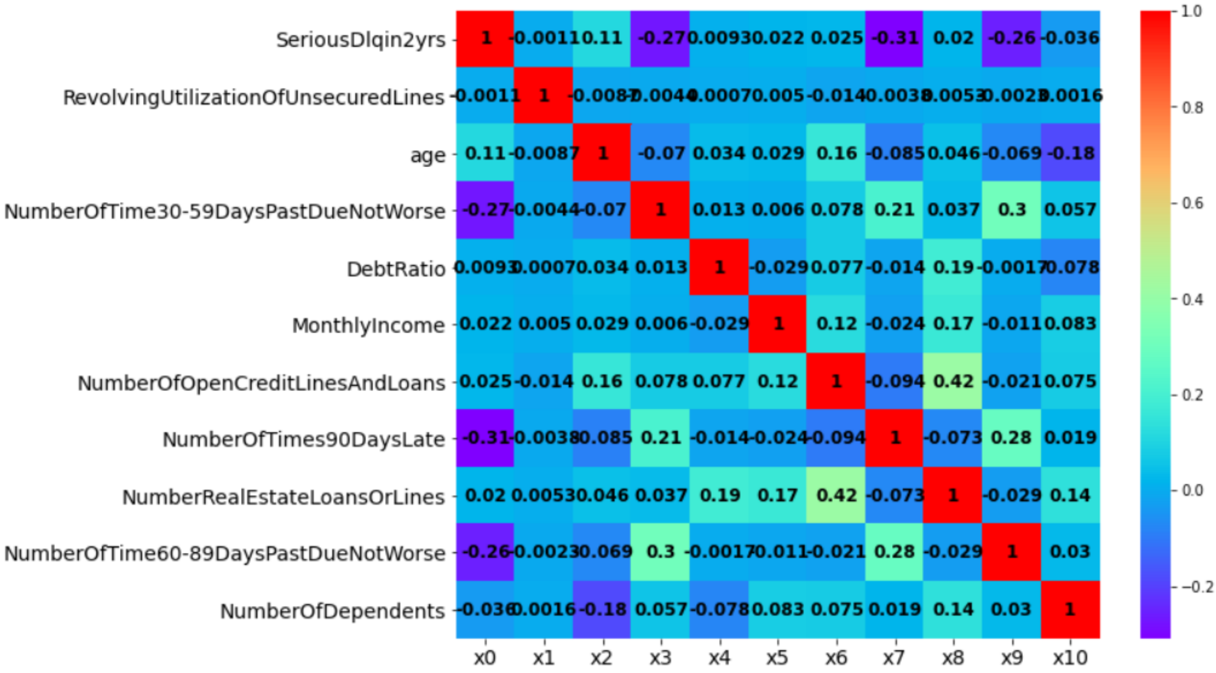


Figure 6. Correlation betwe1en variables

From the Figure 6, we can see that the correlation between variables is very low and the collinearity is not strong, so in order to further explore how different variables explain the dependent variable, IV calculation and evaluation are carried out.

```
Core code:

corr = trainDf.corr()

xticks = ['x0','x1','x2','x3','x4','x5','x6','x7','x8','x9','x10']

yticks = list(corr.index)

fig = plt.figure(figsize=(10,8))
```

```
    ax1 = fig.add_subplot(1, 1, 1)

    sns.heatmap(corr, annot=True, cmap='rainbow',
ax=ax1, annot_kws={'size': 12, 'weight': 'bold',
'color': 'black'})

    ax1.set_xticklabels(xticks, rotation=0,
fontsize=14)

    ax1.set_yticklabels(yticks, rotation=0,
fontsize=14)

    plt.show()
```

7.IV screening

IV, namely Information Value (Information Value), also known as the amount of information. The IV value is used to measure the predictive ability of a variable. The larger the IV value, the stronger the predictive ability of the variable. Usually, in order to ensure the validity of the model and the comprehensiveness of the data, we will provide as many feature variables as possible in the feature engineering, including derivative variables [10]. These derivative variables will not all enter the model for training, otherwise the model will be due to too many related variables. It appears unstable and will increase the complexity of the calculation [3].

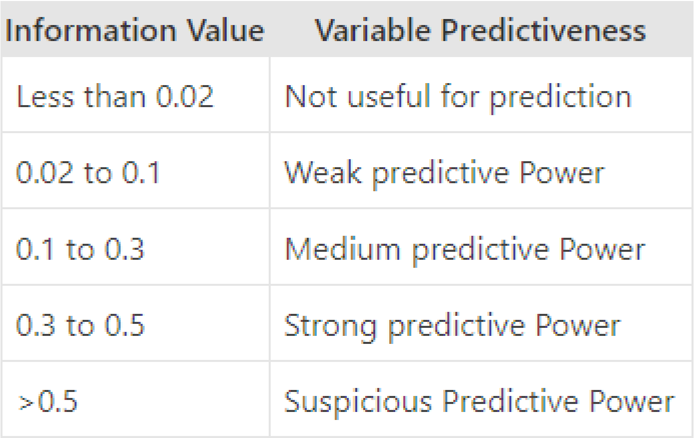
For a grouped variable, the WOE of the i-th group has already been introduced. Similarly, for group i, there will also be a corresponding IV value. The calculation formula is as follows:

$$IV_i = (p_{yi} - p_{ni}) * WOE_i = (p_{yi} - p_{ni}) * ln\left(\frac{p_{yi}}{p_{ni}}\right)$$

The IV value guarantees non-negative results on the basis of WOE. According to the IV value of the variable in each group, the IV value of the entire variable is obtained:

$$IV = \sum_{i=1}^{n} IV_i = \sum_{i=1}^{n} (p_{yi} - p_{ni}) * WOE_i$$

Immediately after binning, the IV of the features is obtained (woe is obtained first, and then IV), which is a value. The formula is: IV=sum((goodattribute-badattribute)*woe), the full name of IV is Infomation Value, Generally used to compare the predictive power of features.



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 7. IV classification chart

According to Siddiqi (2006) and figure 7 shown, by convention the values of the IV statistic in credit scoring can be interpreted as follows [9].

If the IV statistic is:

1. Less than 0.02, then the predictor is not useful for modeling (separating the Goods from the Bads)

2. 0.02 to 0.1, then the predictor has only a weak relationship to the Goods/Bads odds ratio

3. 0.1 to 0.3, then the predictor has a medium strength relationship to the Goods/Bads odds ratio

4. 0.3 to 0.5, then the predictor has a strong relationship to the Goods/Bads odds ratio.

5. > 0.5, suspicious relationship (Check once)

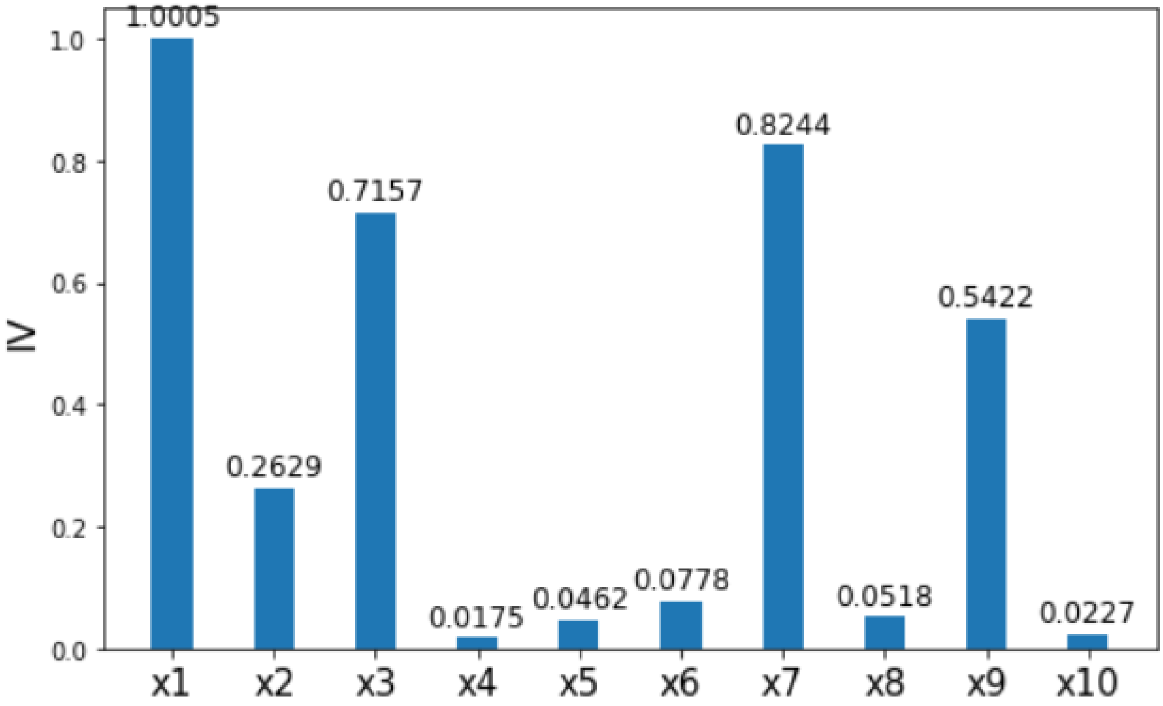


Figure 8. IV diagram

From figure 8, we can see that DebtRatio (x4), MonthlyIncome(x5), NumberOfOpenCreditLinesAndLoans(x6), NumberRealEstateLoansOrLines(x8) and NumberOfDependents(x10) have significantly lower IV values, so they are deleted. We use the remaining variables to build a logistic regression model.

Core code:

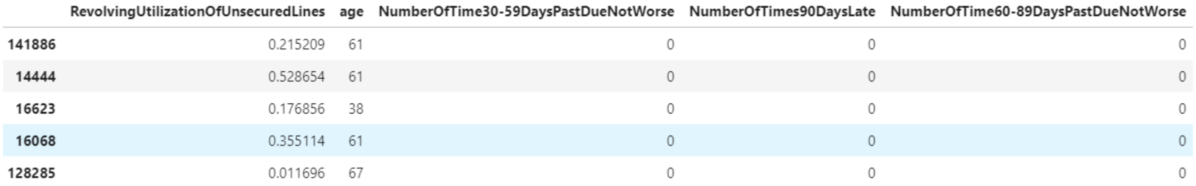
```
ivlist=[ivx1,ivx2,ivx3,ivx4,ivx5,ivx6,ivx7,ivx8,ivx9,ivx10]

index=['x1','x2','x3','x4','x5','x6','x7','x8','x9','x10']
```

```
    fig1 = plt.figure(1,figsize=(8,5))

    ax1 = fig1.add_subplot(1, 1, 1)

    x = np.arange(len(index))+1

    ax1.bar(x,ivlist,width=.4) #
ax1.bar(range(len(index)),ivlist, width=0.4)#generate
hisgram

    #ax1.bar(x,ivlist,width=.04)

    ax1.set_xticks(x)

    ax1.set_xticklabels(index, rotation=0,
fontsize=15)

    ax1.set_ylabel('IV', fontsize=16)
#IV(Information Value),

    #plus numeric on the hisgram

    for a, b in zip(x, ivlist):

        plt.text(a, b + 0.01, '%.4f' % b, ha='center',
va='bottom', fontsize=12)

    plt.show()
```

8.WOE conversion

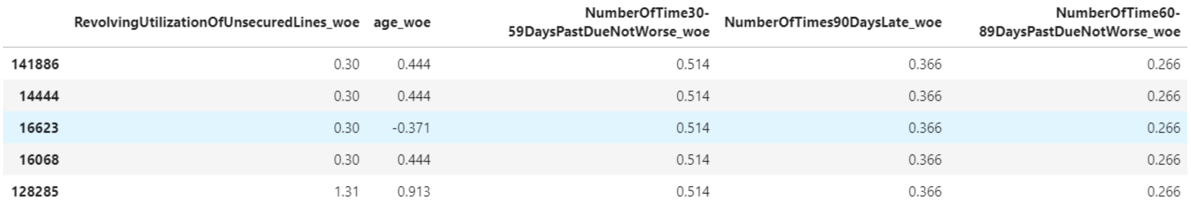
Before building the model, we need to convert the filtered variables into WOE values to facilitate credit scoring

After the transformation, the meaning is more obvious, which can be understood as the difference between the ratio of positive and negative samples in the current group and the ratio of positive and negative samples in all samples. This difference is expressed by the ratio of these two ratios, and then the logarithm. The greater the difference, the greater the WOE, and the greater the probability that the samples in this group will respond. The smaller the difference and the smaller the WOE, the less likely the samples in this group will respond. WOE may be negative, but the greater its absolute value, the greater its contribution to classification. When the ratio of positive and negative in the bin is equal to the ratio of the random (market) positive and negative samples, it means that the bin has no predictive ability, that is, WOE=0.



	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	NumberOfTimes90DaysLate	NumberOfTime60-89DaysPastDueNotWorse
141886	0.215209	61	0	0	0
14444	0.528654	61	0	0	0
16623	0.176856	38	0	0	0
16068	0.355114	61	0	0	0
128285	0.011696	67	0	0	0

Figure 9. The original data of the train dataset



	RevolvingUtilizationOfUnsecuredLines_woe	age_woe	NumberOfTime30-59DaysPastDueNotWorse_woe	NumberOfTimes90DaysLate_woe	NumberOfTime60-89DaysPastDueNotWorse_woe
141886	0.30	0.444	0.514	0.366	0.266
14444	0.30	0.444	0.514	0.366	0.266
16623	0.30	-0.371	0.514	0.366	0.266
16068	0.30	0.444	0.514	0.366	0.266
128285	1.31	0.913	0.514	0.366	0.266

Figure 10. Replace the variable with the value of the woe function

Figures 9 and 10 are the original data and the data after WOE replacement. From Figures 9 and 10, it can be seen that different attributes correspond to a WOE value in a certain interval, which is the establishment of a logistic regression model in the future. The scorecard is built to create the foundation.

```
Core code:

def trans_woe(var,var_name,woe,cut):

    woe_name=var_name+'_woe'

    for i in range(len(woe)):

# len(woe) Get how many values are in woe

        if i==0:
```

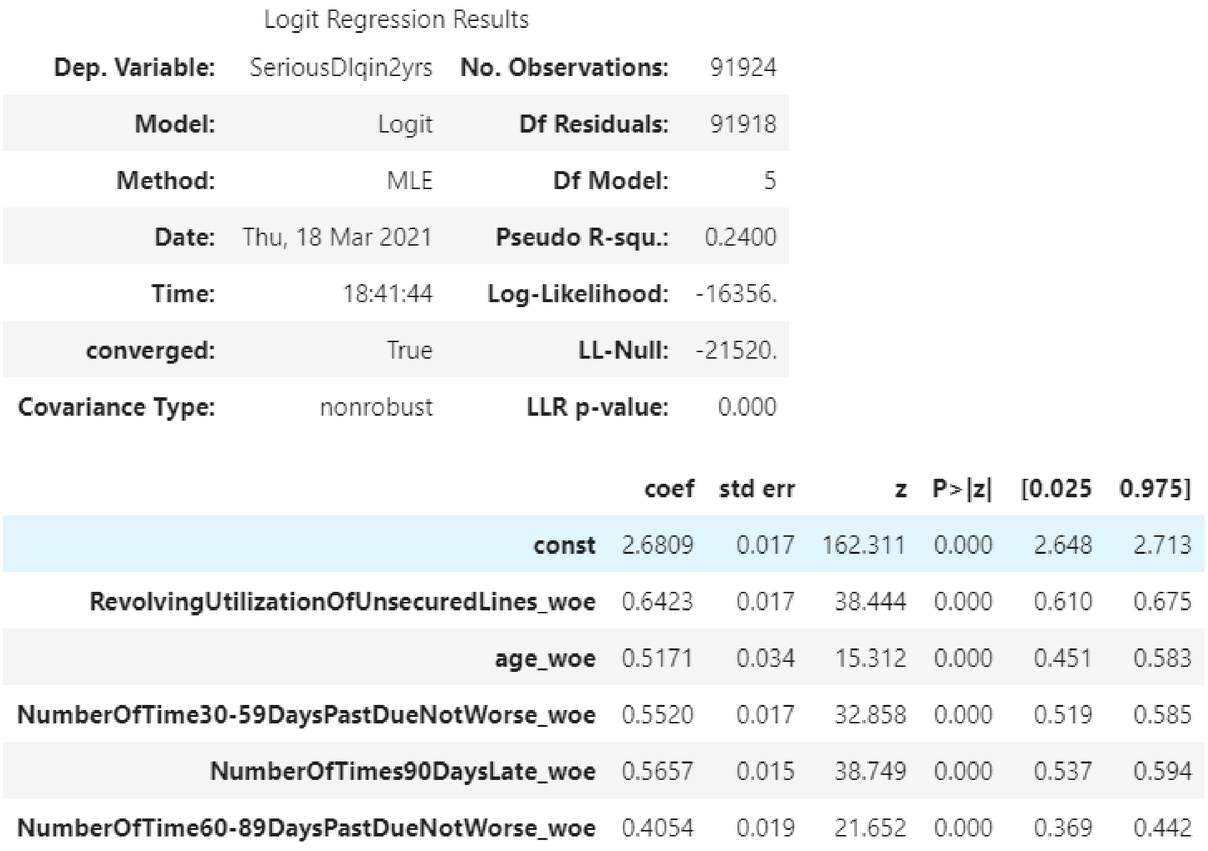
```
var.loc[(var[var_name]<=cut[i+1]),woe_name]=woe[i]
    #The value of woe is assigned to the woe_name
column of var according to the lower node of the cut
bin, the first paragraph of the bin
            elif (i>0) and  (i<=len(woe)-2):

var.loc[((var[var_name]>cut[i])&(var[var_name]<=cut[i
+1])),woe_name]=woe[i]


            else:
                var.loc[(var[var_name]>cut[len(woe)-
1]),woe_name]=woe[len(woe)-1]
    #Greater than the upper limit of the last binning
interval, the last value is positive infinity
    return var


     X_org=trainDf.loc[:,['RevolvingUtilizationOfUnse
curedLines','age','NumberOfTime30-
59DaysPastDueNotWorse','NumberOfTimes90DaysLate','Num
berOfTime60-89DaysPastDueNotWorse']]
    X_org.head()


    x1_name='RevolvingUtilizationOfUnsecuredLines'
    x2_name='age'
    x3_name='NumberOfTime30-59DaysPastDueNotWorse'
    x7_name='NumberOfTimes90DaysLate'
    x9_name='NumberOfTime60-89DaysPastDueNotWorse'

    trainDf=trans_woe(trainDf,x1_name,woex1,cutx1)
    trainDf=trans_woe(trainDf,x2_name,woex2,cutx2)
    trainDf=trans_woe(trainDf,x3_name,woex3,cutx3)
    trainDf=trans_woe(trainDf,x7_name,woex7,cutx7)
    trainDf=trans_woe(trainDf,x9_name,woex9,cutx9)

    Y=trainDf['SeriousDlqin2yrs']
    #Independent variables, eliminate variables that
have no obvious impact on the dependent variable
    X=trainDf.drop(['SeriousDlqin2yrs','DebtRatio','M
onthlyIncome',
'NumberOfOpenCreditLinesAndLoans','NumberRealEstateLo
ansOrLines','NumberOfDependents'],axis=1)
    X=trainDf.iloc[:,-5:]
    X.head()
```

9.Machine learning model

A variety of machine learning models can be used to predict user behavior, but for the subsequent establishment of user credit score cards, a logistic regression model is used. The logistic regression model can be directly converted into a user score card [6]. At the same time, the accuracy of the logistic regression model on the binary classification problem is higher than that of other machine learning models [11].

9.1 Logistic regression model establishment

In this project, logistic regression is used to establish a machine learning model. Coefficient in the logistic regression model is of great significance for the establishment of credit score cards. Figure 11 is the relevant information of the logistic regression model.



Logit Regression Results

Dep. Variable:	SeriousDlqin2yrs	**No. Observations:**	91924
Model:	Logit	**Df Residuals:**	91918
Method:	MLE	**Df Model:**	5
Date:	Thu, 18 Mar 2021	**Pseudo R-squ.:**	0.2400
Time:	18:41:44	**Log-Likelihood:**	-16356.
converged:	True	**LL-Null:**	-21520.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.6809	0.017	162.311	0.000	2.648	2.713
RevolvingUtilizationOfUnsecuredLines_woe	0.6423	0.017	38.444	0.000	0.610	0.675
age_woe	0.5171	0.034	15.312	0.000	0.451	0.583
NumberOfTime30-59DaysPastDueNotWorse_woe	0.5520	0.017	32.858	0.000	0.519	0.585
NumberOfTimes90DaysLate_woe	0.5657	0.015	38.749	0.000	0.537	0.594
NumberOfTime60-89DaysPastDueNotWorse_woe	0.4054	0.019	21.652	0.000	0.369	0.442

Figure 11. Logistic regression model

```
Core code:

X1=sm.add_constant(X)

logit=sm.Logit(Y,X1)

result=logit.fit()

result.summary()
```

9.2 Logistic regression model accuracy test

Model evaluation: We need to verify the predictive ability of the model. We use the test data reserved at the beginning of the modeling phase for verification.

Evaluate the fitting ability of the model through ROC curve and AUC. ROC is Receiver Operating Characteristic the area of the ROC curve is AUC (Area Under the Curve), it is used to measure the performance of the machine learning algorithm for the "two classification problem" (generalization ability). The total number of true positive categories in the sample is TP+FN. TPR is True Positive Rate, TPR = TP/(TP+FN). Similarly, the total number of true counterexample categories in the sample is FP+TN. FPR is False Positive Rate, FPR=FP/(TN+FP) [13].

In Python, you can use sklearn.metrics, which can easily compare two classifiers and automatically calculate ROC and AUC.

Draw the FPR and TPR in the result into two-dimensional coordinates as figure 11 shown. The ROC curve obtained is as follows (indicated by the blue line), and the area of the ROC curve is represented by AUC (the area under the blue curve).

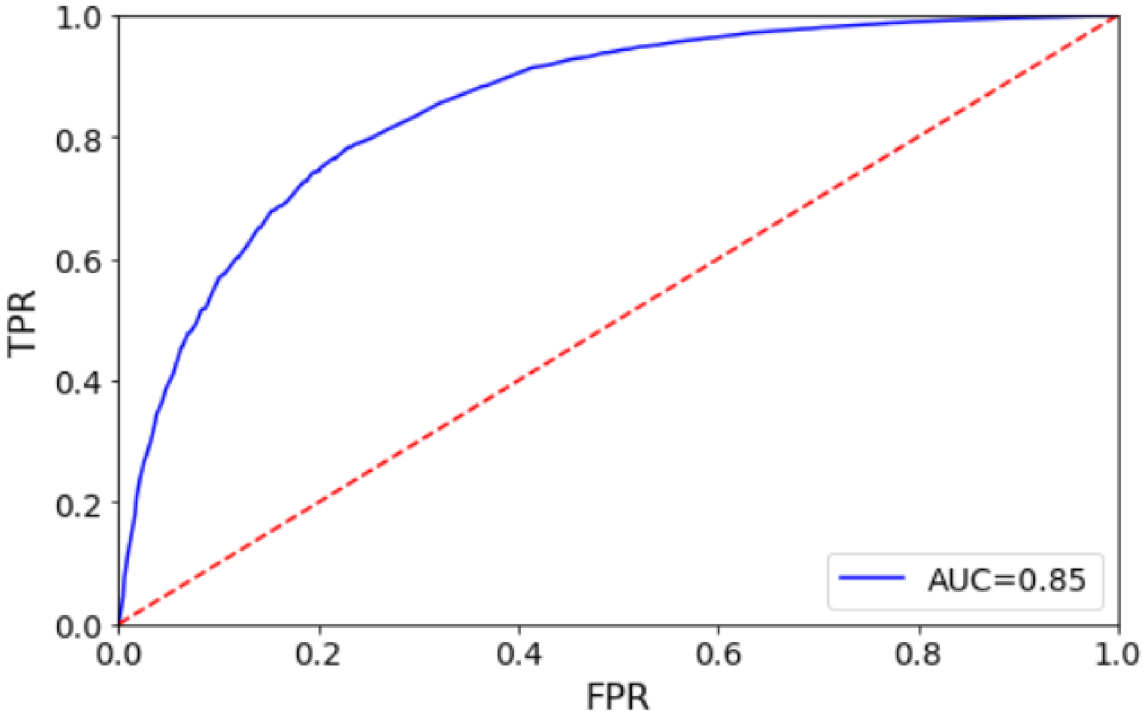


Figure 12. ROC curve and AUC value of the model

As shown in Figure 12, an AUC of 0.85 indicates that the accuracy of the model is 85% or relatively accurate.

```
    Core code:

    testDf=trans_woe(testDf,x1_name,woex1,cutx1)

    testDf=trans_woe(testDf,x2_name,woex2,cutx2)

    testDf=trans_woe(testDf,x3_name,woex3,cutx3)

    testDf=trans_woe(testDf,x7_name,woex7,cutx7)

    testDf=trans_woe(testDf,x9_name,woex9,cutx9)

    #Building the characteristics and labels of the
test set

    test_X=testDf.iloc[:,-5:]   #Test data
characteristics

    test_Y=testDf.iloc[:,0]     #Test Data Label
```

```
#Assessment
from sklearn import metrics
X3=sm.add_constant(test_X)
resu = result.predict(X3)
fpr,tpr,threshold=metrics.roc_curve(test_Y,resu)
rocauc=metrics.auc(fpr,tpr)   #calculate AUC

#Should plus a title
plt.figure(figsize=(8,5))
plt.plot(fpr,tpr,'b',label='AUC=%0.2f'% rocauc)
plt.legend(loc='lower right',fontsize=14)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.ylabel('TPR',fontsize=16)
plt.xlabel('FPR',fontsize=16)
plt.show()
```

10.Credit score card

Less well known but equally important are credit and behavioral scoring, which are the applications of financial risk forecasting to consumer lending. An adult in the UK or US is being credit scored or behaviors scored on average at least once a week as the annual reports of the credit bureau imply. The fact that most people are not aware of being scored does not diminish from its importance. This area of financial risk has a limited literature with only a few surveys (Rosenberg & Gleit, 1994, Hand & Henley, 1997, Thomas, 1992, Thomas, 1998) and a handful of books (Hand & Jacka, 1998, Thomas et al., 1992, Lewis, 1992, Mays, 1998). The aim of this survey is to give an overview of the objectives, techniques and difficulties of credit scoring as an application of forecasting [7].

10.1 Credit score card Creation

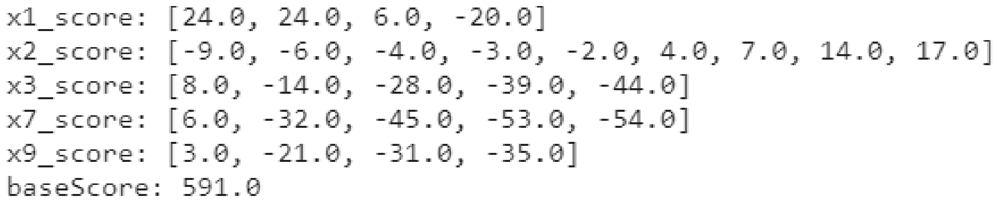
The format of the standard scorecard is that each variable in the scorecard follows a series of IF-THEN rules. The value of the variable determines the value of the variable assigned, and the total score is the sum of the scores of each variable [5].

Before establishing a standard scorecard, we need to select several scorecard parameters: basic score, PDO (score for doubling the ratio), and good to bad ratio. Here, we take 600 as the basic score, PDO as 20 (the ratio of good to bad is doubled for every 20 points higher), and the ratio of good to bad is 20.

Basic_score is 600, PDO is 20 and Coef_const is 2.6809 according the figure 10 shown. To get basic line for credit card, To get basic points, we should get Factor and Offset first. Factor = PDO /log(2) and Offset = Basic_score – PDO*log(PDO)/log(2), after getting the Factor and Offset, according to the formula of Base_line: Base_line = Offset + Factor*Coef_const , substituting all the data into the formula, you can get the score card of age as the figure 12 shown at the last row in the figure, you can calculate the Base_line as 591.

Due to the many parameters that need to be calculated, this project uses age as an example to show the process of establishing a credit score card for this variable.

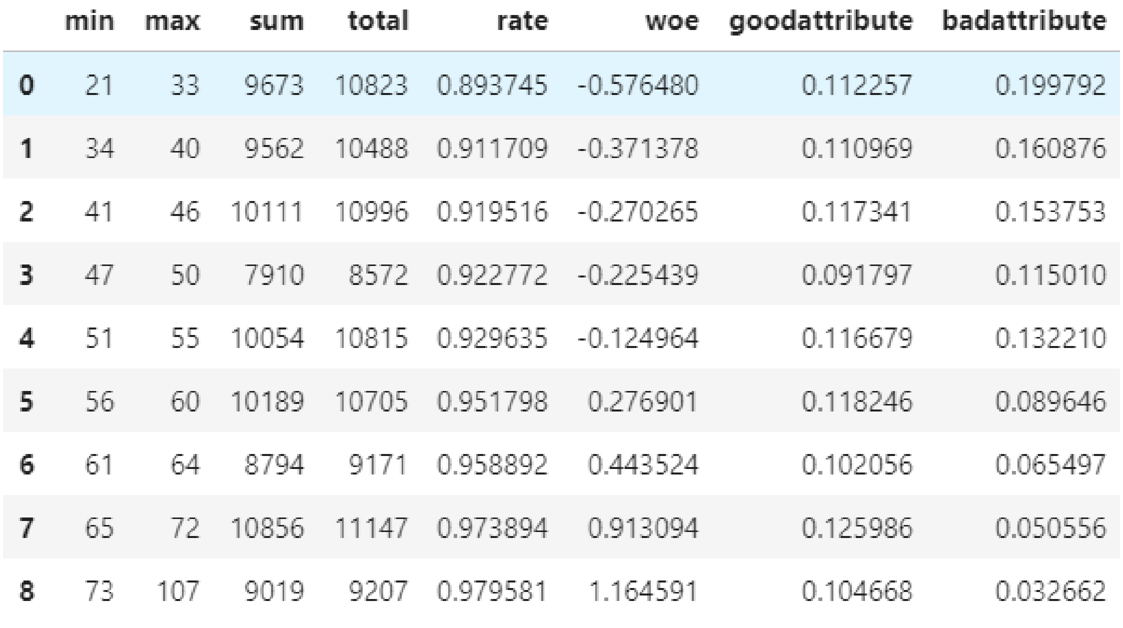
In order to get the corresponding score of each segment of age, PDO, coefficient_age and woe_age are needed. We can get coefficient_age in the logistic regression model as the figure 10 shown in the third row of the figure, and we can get woe_age in the woe calculation conversion of age as the figure 13 shown. The age Offset = Coef_age * Woe_age * Factor. Substituting all the data into the formula, you can get the score card of age as the table 2 shown.



```
x1_score: [24.0, 24.0, 6.0, -20.0]
x2_score: [-9.0, -6.0, -4.0, -3.0, -2.0, 4.0, 7.0, 14.0, 17.0]
x3_score: [8.0, -14.0, -28.0, -39.0, -44.0]
x7_score: [6.0, -32.0, -45.0, -53.0, -54.0]
x9_score: [3.0, -21.0, -31.0, -35.0]
baseScore: 591.0
```

Figure 13. Score for each interval of different attributes of the scorecard

According to the WOE info as figure 13 shown of every variables, can get every interval's score and can make a credit score card for every estimated variables.



	min	max	sum	total	rate	woe	goodattribute	badattribute
0	21	33	9673	10823	0.893745	-0.576480	0.112257	0.199792
1	34	40	9562	10488	0.911709	-0.371378	0.110969	0.160876
2	41	46	10111	10996	0.919516	-0.270265	0.117341	0.153753
3	47	50	7910	8572	0.922772	-0.225439	0.091797	0.115010
4	51	55	10054	10815	0.929635	-0.124964	0.116679	0.132210
5	56	60	10189	10705	0.951798	0.276901	0.118246	0.089646
6	61	64	8794	9171	0.958892	0.443524	0.102056	0.065497
7	65	72	10856	11147	0.973894	0.913094	0.125986	0.050556
8	73	107	9019	9207	0.979581	1.164591	0.104668	0.032662

Figure 14. Example of age attribute and its info

Variable	Interval	Score
Age	(21,34]	-9
	(34,41]	-6
	(41,47]	-4
	(47,51]	-3
	(51,56]	-2
	(56,61]	4
	(61,65]	7
	(65,73]	14
	(73,107]	17

Table 2. Age credit score card

According to Figure 14 and Table 2 we can see that as the age increases, the woe value also increases, and the corresponding score also increases. The higher the woe value, the smaller the badattribute. It can be seen that the older the age, the higher the credit rating.

The paper points out that the statistical scoring models discussed in the literature have focused primarily on the minimization of default rates, which is in fact only one dimension of the more general problem of granting credit. To the extent that for the lender profit maximization or cost minimization is, or should be, the objective of a scoring model, then most of the applied literature seems incomplete [8].

Then according to the establishment process of the age score card, the selected attributes are also established for the score card, and finally aggregated into the total bank customer credit score card as table 3 shown.

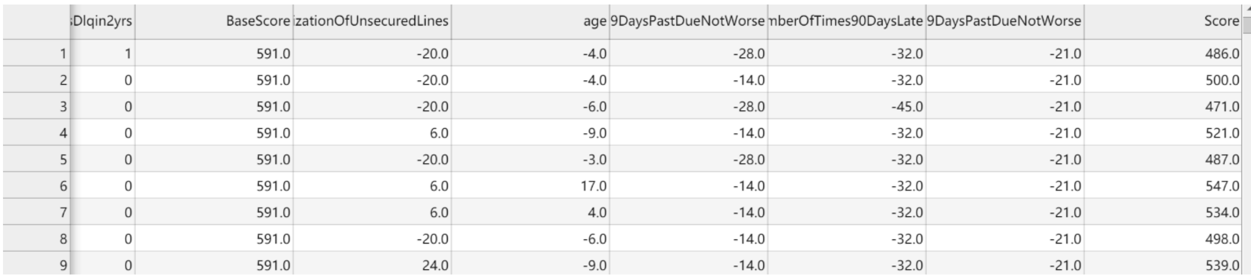
Variable	Interval	Score
Base_line	---	591
RevolvingUtilizationOfUnsecuredLines	<=0.289	24
	(0.289,0.348]	24
	(0.348,0.543]	6
	>0.543	-20
Age	<=34	-9
	(34,41]	-6
	(41,47]	-4
	(47,51]	-3
	(51,56]	-2
	(56,61]	4
	(61,65]	7
	(65,73]	14
	>73	17
NumberOfTime30-59DaysPastDueNotWorse	0	8
	1	-14
	[2,3]	-28
	[4,5]	-39
	[6,12]	-44
NumberOfTimes90DaysLate	0	6
	1	-32
	[2,3]	-45
	[4,5]	-53
	[6,13]	-54
NumberOfTime60-89DaysPastDueNotWorse	0	3
	1	-21
	[2,3]	-31
	[4,11]	-35

Table 3. Credit score card

According to Table 3, We can see that only five variables were used to construct the scorecard in the end. Because the IV values of the remaining variables were too low and did not have good predictive ability, these variables with low IV values were eliminated. Excluding variables with low IV values will not affect the accuracy of the model and the accuracy of the scorecard. Adding unnecessary variables will have a greater impact on the model.

The calculation formula for the final credit score of bank customers can be obtained:

$$Bankcustomercreditscore = Base\_line\ score + Sum(various\_attribute\ scores)$$



	;Dlqin2yrs	BaseScore	izationOfUnsecuredLines	age	9DaysPastDueNotWorse	nberOfTimes90DaysLate	9DaysPastDueNotWorse	Score
1	1	591.0	-20.0	-4.0	-28.0	-32.0	-21.0	486.0
2	0	591.0	-20.0	-4.0	-14.0	-32.0	-21.0	500.0
3	0	591.0	-20.0	-6.0	-28.0	-45.0	-21.0	471.0
4	0	591.0	6.0	-9.0	-14.0	-32.0	-21.0	521.0
5	0	591.0	-20.0	-3.0	-28.0	-32.0	-21.0	487.0
6	0	591.0	6.0	17.0	-14.0	-32.0	-21.0	547.0
7	0	591.0	6.0	4.0	-14.0	-32.0	-21.0	534.0
8	0	591.0	-20.0	-6.0	-14.0	-32.0	-21.0	498.0
9	0	591.0	24.0	-9.0	-14.0	-32.0	-21.0	539.0

Figure 15. Bank users based on the score card score

Substituting all the scorecard data into the original data, a new customer table can be obtained as shown in figure 15. The different information of the customer is quantitatively analyzed, and the customer's total credit score is obtained.

```
    Core code:

    p=20/np.log(2)#Scale Factor

    q=600-20*np.log(20)/np.log(2)#Equal to offset, 
offset

    x_coe=[2.6809,0.6423,0.5171,0.5520,0.5657,0.4054]
#Regression coefficients

    baseScore=round(q+p*x_coe[0],0)

    #Personal total score = basic score + each part 
score


    def get_score(coe,woe,factor):

        scores=[]

        for w in woe:

            score=round(coe*w*factor,0)

            scores.append(score)
        return scores
    #Each item score
    x1_score=get_score(x_coe[1],woex1,p)
    x2_score=get_score(x_coe[2],woex2,p)
    x3_score=get_score(x_coe[3],woex3,p)
    x7_score=get_score(x_coe[4],woex7,p)
    x9_score=get_score(x_coe[5],woex9,p)

    def compute_score(series,cut,score):
        list = []
        i = 0
        while i < len(series):
            value = series[i]
            j = len(cut) - 2
            m = len(cut) - 2
            while j >= 0:
                if value >= cut[j]:
                    j = -1
                else:
                    j -= 1
                    m -= 1
            list.append(score[m])
            i += 1
        return list
    test1 = pd.read_csv("cs-training.csv")
    test1['BaseScore']=np.zeros(len(test1))+baseScore
    test1['x1']
=compute_score(test1['RevolvingUtilizationOfUnsecured
Lines'], cutx1, x1_score)
    test1['x2'] = compute_score(test1['age'], cutx2,
x2_score)
    test1['x3'] =
compute_score(test1['NumberOfTime30-
59DaysPastDueNotWorse'], cutx3, x3_score)
    test1['x7'] =
compute_score(test1['NumberOfTimes90DaysLate'],  cutx7,
x7_score)
    test1['x9'] =
compute_score(test1['NumberOfTime60-
89DaysPastDueNotWorse'],cutx9,x9_score)
    test1['Score'] = test1['x1'] + test1['x2'] +
test1['x3'] + test1['x7'] +test1['x9']  + baseScore


    scoretable1=test1.iloc[:,[1,-7,-6,-5,-4,-3,-2,-1]]
#Select the required column, which is the rating
column
    scoretable1.head()
    scoretable1.to_csv('ScoreData_simple_version.csv')
    colNameDict={'x1': 'RevolvingUtilizationOfUnsecuredLines' ,'x2':'age','x3':'NumberOfTime30-59DaysPastDueNotWorse',
                 'x7':'NumberOfTimes90DaysLate', 'x9':'NumberOfTime60-89DaysPastDueNotWorse'}
    scoretable2=scoretable1.rename(columns=colNameDict,inplace=False)
    scoretable2.to_csv('ScoreData.csv')


    p = 20/np.log(2)
    q = 600 - 20*np.log(20)/np.log(2)


    def get_score(coe,woe,factor):
        scores=[]
        for w in woe:
            score=round(coe*w*factor,0)
            scores.append(score)
        return scores
    x_coe = [2.6809,0.6423,0.5171,0.5520,0.5657,0.4054]
    baseScore = round(q + p * x_coe[0], 0)


    x1_score=get_score(x_coe[1],woex1,p)
    x2_score=get_score(x_coe[2],woex2,p)
```

```
x3_score=get_score(x_coe[3],woex3,p)
x7_score=get_score(x_coe[4],woex7,p)
x9_score=get_score(x_coe[5],woex9,p)


print('x1_score:',x1_score)
print('x2_score:',x2_score)
print('x3_score:',x3_score)
print('x7_score:',x7_score)
print('x9_score:',x9_score)
print('baseScore:',baseScore)
```

10.2 Credit Score Card Accuracy Verification

After the bank user credit score card is built, the original data is converted into a credit score, and the user's credit situation is quantitatively evaluated. However, whether different credit scores can reflect different credit conditions requires further exploration and analysis. Therefore, this project uses the following methods for evaluation.

Divide new dataset into 16 groups and calculate the bad customer rate of every group, the bad customer rate is bad customer in the group divide all customer in the group and then compare all groups bad customer rate.

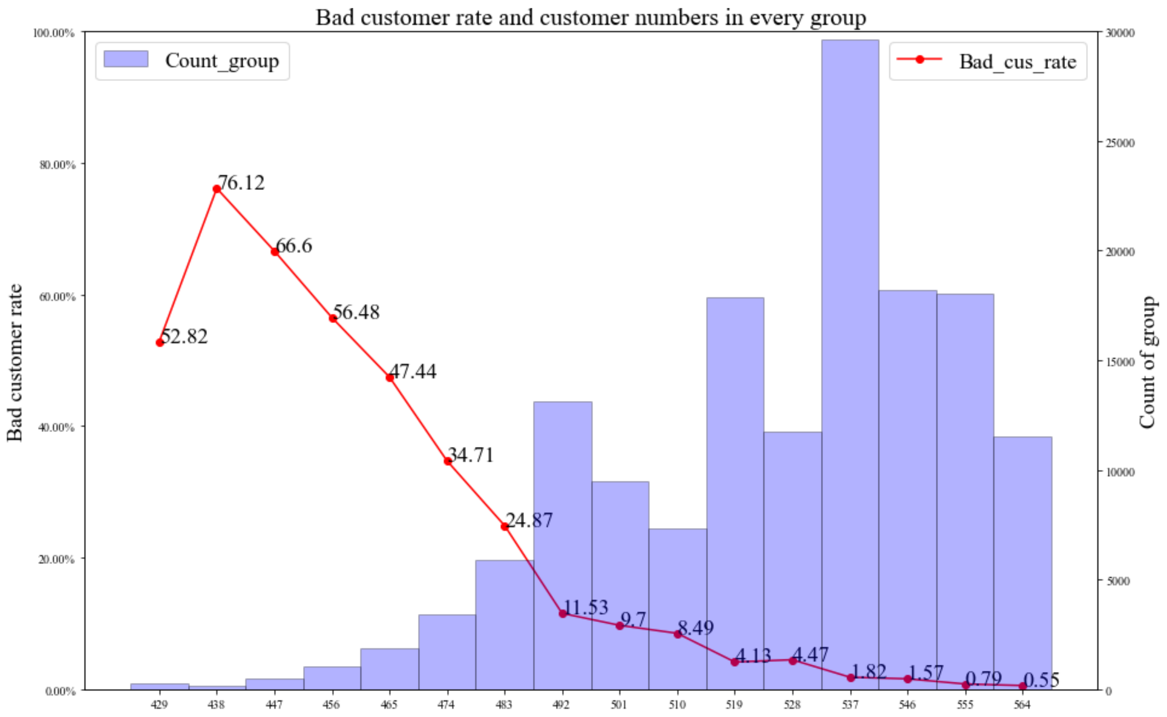


Figure 16. The impact of different groups of customer credit scores on customer default rates

From figure 16, This graph is more complicated, consisting of X axis, left Y axis, and right Y axis, the X axis is different credit scores, the left Y axis is the user's default probability, and the right Y axis is the number of users, we can see that as the customer score increase the bad customer rate decrease, although there are some fluctuations in the middle, the overall trend is that as the customer's credit score increases, the user's default rate decreases, we can get conclusion that the credit score card can measure the user's creditworthiness. Especially when the customer's credit score is greater than 500, the customer's default probability drops to less than 10%, which provides a good reference basis for customer segmentation and rating.

Core code:

```
group_cus_count_lis = []

group_bad_cus_count_lis = []

group_bad_rate_lis = []
```

```
    for i in range(429,565,9):
        group_cus_count =
data_cred.loc[(data_cred['Score']>=i) &
(data_cred['Score']<= i+8)].shape[0]
    #     print(group_cus_count)
        group_cus_count_lis.append(group_cus_count)
        group_bad_cus_count =
data_cred.loc[(data_cred['Score']>=i) &
(data_cred['Score']<= i+8) &
(data_cred['SeriousDlqin2yrs']==1)].shape[0]
    #     print(group_bad_cus_count)



group_bad_cus_count_lis.append(group_bad_cus_count)
        if group_cus_count != 0:
            group_bad_rate =
group_bad_cus_count/group_cus_count*100
            group_bad_rate = round(group_bad_rate, 2)
    #         group_bad_rate = format(group_bad_rate,
'.2%')
            group_bad_rate_lis.append(group_bad_rate)
        else:
            group_bad_rate_lis.append('None')
    a = group_cus_count_lis
    b= group_bad_rate_lis
    l=[i for i in range(429,565,9)]


    plt.rcParams['font.sans-serif']=['Times New
Roman']


    fmt='%.2f%%'
    yticks = mtick.FormatStrFormatter(fmt)
    lx=[u'429',u'438',u'447',u'456',u'465',u'474',u'4
83',u'492',u'501',u'510',u'519',u'528',u'537',u'546',
u'555',u'564']


    font1 = {'family' : 'Times New Roman',
    'weight' : 'normal',
    'size'   : 18,
    }


    fig = plt.figure(figsize=(15, 10))


    ax1 = fig.add_subplot(111)
    ax1.plot(l, b,'or-',label=u'Bad_cus_rate');
    ax1.yaxis.set_major_formatter(yticks)
    for i,(_x,_y) in enumerate(zip(l,b)):

plt.text(_x,_y,b[i],color='black',fontsize=18,)
```

```
    ax1.legend(loc=1,fontsize = 18)

    ax1.set_ylim([0, 100]);

    ax1.set_ylabel('Bad customer rate',fontsize = 18);

    plt.legend(prop={'family':'Times New
Roman','size':18})

    ax2 = ax1.twinx() # this is the important
function

    plt.bar(l,a,alpha=0.3,color='blue',label=u'Count_
group', width=9, edgecolor='black')

    # plt.bar(np.arange(16), y, alpha=0.5, width=0.3,
color='yellow', edgecolor='red', label='The First
Bar', lw=3)

    ax2.legend(loc=2,fontsize = 18)

    ax2.set_ylim([0, 30000])

    ax2.set_ylabel('Count of group',fontsize = 18);

    plt.legend(prop=font1,loc="upper left")

    plt.xticks(l,lx,size = 18)

    plt.title('Bad customer rate and customer numbers
in every group',fontsize = 20)

    plt.show()
```

10.3Credit Score Card Classification

According to the user's credit score, the customer is classified to provide a simpler basis for whether the bank accepts the user's loan application [4]. In order to solve this problem, Figure 17 is drawn.

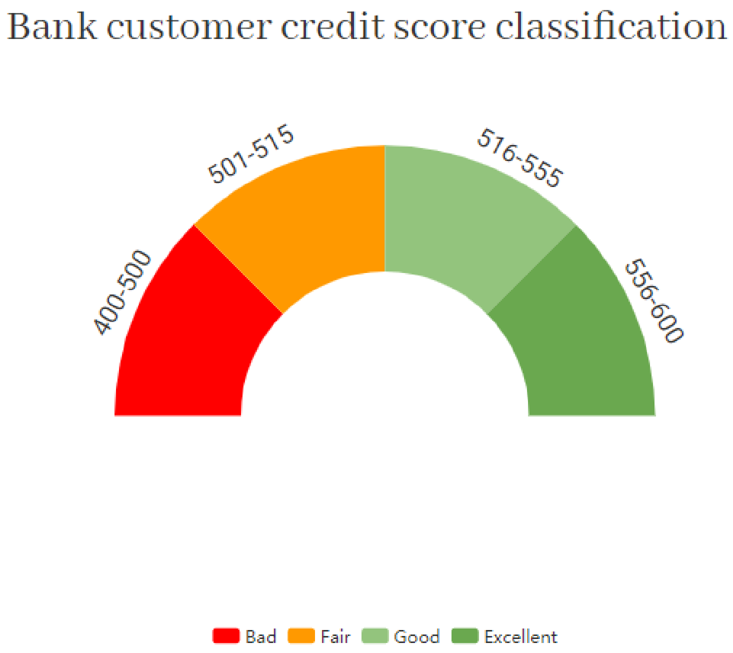


Figure 17. Bank customer credit rating classification

According to Figure 16, according to the bank users' credit scores and the default probability, users are divided into four categories, bad, fair, good, and excellent. Based on this information, Figure 17 is obtained. Bad customer’s credit score is 400-500, and the default probability is more than 10%. Banks will not accept loan applications from these users. Fair customer’s credit score is 501-515, and the default probability is less than 10% and more than 5%. The bank accepts loan applications from these users, but their loan amount is average. Good customers have a credit score of 516-555, and their default probability is less than 5%, and more than 1%, their loan amount will be better than a Fair customer. Excellent customers have a credit score of 556-600, and their default probability is 1% Below, their loan amount is the highest.

In short, the higher the user's credit score, the easier it is for their loan application to be accepted, and the higher their loan amount. However, how to specify the minimum score for loan acceptance and maximize the profit of the bank's loan business requires a more in-depth analysis by the bank, as well as historical inspection.

11. Conclusion:

1. Logistic regression is a powerful model for predicting the prediction of binary classification results. The logistic regression model is especially suitable for the establishment of credit score cards.

2. The establishment of a machine learning model should propose strong sharing variables, which is beneficial to improve the accuracy and interpretability of the model, and the heat map can check the sharing between variables. It is not as good as more independent variables to build a model. It is necessary to eliminate multiple collinearity and irrelevant variables.

3. IV is a good reference standard for judging the predictive ability of variables. You can decide which variables to use to build a model based on the IV value.

4. Credit score card is a good method for qualitative and quantitative analysis of users, and can be continuously revised in future practice. Banks can set the lower limit of acceptance of loan customers' credit scores according to their own business risk preferences. A good model is the model that maximizes the benefits of the bank's loan business, not the least risky model.

12.Social responsibility

12.1 Introduction

The developed project aims to use machine learning and data mining to analyze bank user information and establish user credit score cards to provide a reference for the bank's loan business. The development of the program is only carried out with the help of computer.

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

The work was carried out in the hall of residence of TPU (8th floor). Room 806B was a research execution place.

The layout of the room is shown in Figure 18

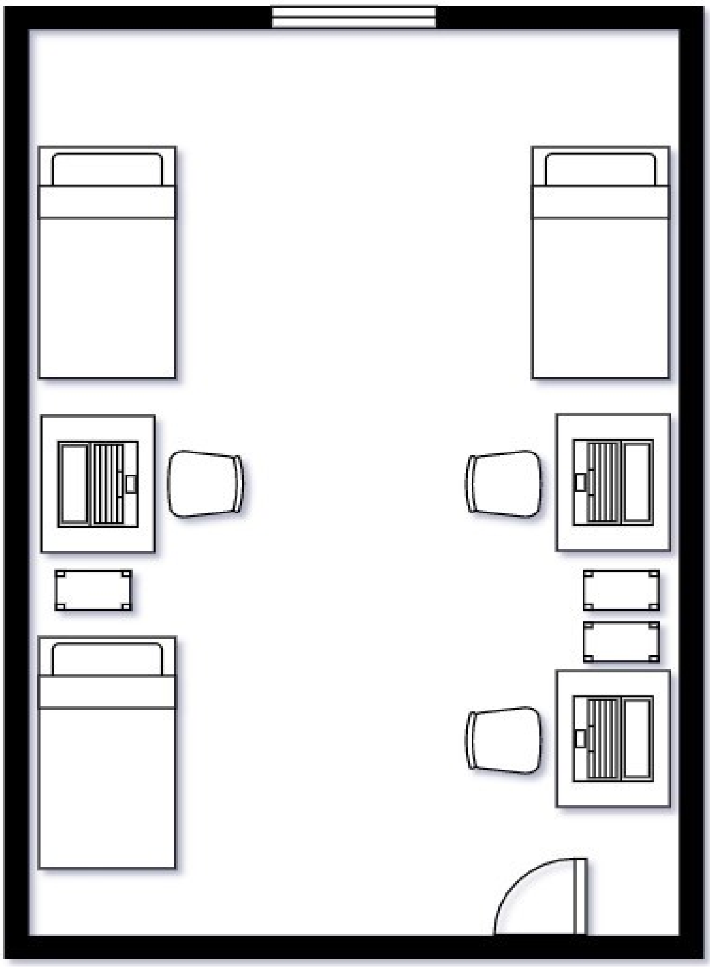


Figure 18. Room layout 806B

12.2 Legal and organizational issues of occupational safety

Nowadays one of the main ways to radical improvement of all prophylactic work referred to reduce Total Incidents Rate and occupational morbidity is the widespread implementation of an integrated Occupational Safety and Health management system. That means combining isolated activities into a single 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 [14], every employee has the right:

- To have a workplace that meets Occupational safety requirements;
- To have a compulsory social insurance against accidents at manufacturing and 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 r11egulatory 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.

12.3 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 [15].

12.4 Occupational safety

Workplace safety is the responsibility of everyone in the organization.

Occupational hygiene is a system of ensuring the health of workers in the process of labor activity, including legal, socio-economic, organizational and technical, sanitary and hygienic, treatment and prophylactic, rehabilitation and 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 processthat can cause occupational pathology, temporary or permanent decrease in working capacity, increase the frequency of somatic and infectious diseases, and lead to impaired health of the offspring.

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

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

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

Table 4 - Potential hazardous and harmful production factors

Factors **(GOST 12.0.003-2015)**	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Increased levels of noise	+	+		**GOST 12.1.003-2014** Occupational safety standards system. Noise. General safety requirements
2. Lack or lack of natural light, insufficient illumination	+			**SanPiN 2.2.1/2.1.1.1278-03** Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings
3. Electromag neticfields	+	+	+	**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.

Increased levels of noise

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

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

Lack or lack of natural light, insufficient illumination

Light sources can be both natural and artificial. The natural source of the light in the room is the sun, artificial light are lamps. 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 [17].

According to the SanPiN 2.2.1/2.1.1.1278-03 [17] standard., the illumination
on the table surface in the area of the working document should be 300-500 lux. Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300 lux.

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

Electromagnetic fields

In this case, the sources of increased intensity of the electromagnetic field are a personal computer. 8 kA / m is considered acceptable. An hour's working day for an employee at his workplace, with the maximum permissible level of tension, should be no more than 8 kA / m, and the level of magnetic induction should be 10 mT [18]. 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 [18].

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 [18]:

- 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 [21], 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.

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) [19].

12.4 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 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 [22].

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 [22].

12.5 Safety in emergency

An emergency situation (ES) is a situation in a certain territory that has developed as a result of an accident, hazardous natural phenomenon, catastrophe or other disaster, which may entail human casualties, damage to human health or the environment, significant material losses and violation of the living conditions of people. Emergency for the presented work space is a fire. This emergency can occur in the event of non-compliance with fire safety measures, violation of the technique of using electrical devices and PCs, violations of the wiring of electrical networks and a number of other reasons.

The working space provided for the performance of the WRC, according to SanPiN 2.2.1 / 2.1.1.1278-03[4], can be classified as category B (fire hazard).

The following reasons can be indicated as possible causes of a fire:

- short circuit.
- dangerous overload of networks, which leads to strong heating of live

parts and ignition of insulation.11

- start-up of equipment after incorrect and unqualified repairs.

To prevent emergencies, it is necessary to comply with fire safety rules in order to ensure the state of protection of employees and property from fire

To protect against short circuits and overloads, it is necessary to correctly select, install and use electrical networks and automation equipment.

To prevent the occurrence of fires, it is necessary to exclude the formation of a combustible environment, to monitor the use of non-combustible or hardly combustible materials in the construction and decoration of buildings.

It is necessary to carry out the following fire prevention measures:

- organizational measures related to the technical process, taking into account the fire safety of the facility (personnel briefing, training in safety rules, publication of instructions, posters, evacuation plans).
- operational measures that consider the operation of the equipment used (compliance with equipment operating standards, ensuring a free approach to equipment, maintaining conductor insulation in good condition).
- technical and constructive measures related to the correct placement and installation of electrical equipment and heating devices (compliance with fire safety measures when installing electrical wiring, equipment, heating, ventilation and lighting systems).

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

An evacuation plan is presented in the presented working room at the entrance, a fire alarm system is installed. The room is equipped with OU-2 type carbon dioxide fire extinguishers in the amount of 2 pieces per one working area. There is an electrical panel within the reach of workers, with the help of which it is possible to completely de-energize the working room.

In the event of a fire, you must call the fire department by phone 101 and inform the place of the emergency, take measures to evacuate workers in accordance with the evacuation plan. In the absence of direct threats to health and life, make an attempt to extinguish the resulting fire with existing carbon dioxide fire extinguishers. In case of loss of control over the fire, it is necessary to evacuate after the employees according to the evacuation plan and wait for the arrival of the fire service specialists.

12.6 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.

13. Financial management, resource efficiency and resource saving

The purpose of this section is to discuss the issues of competitiveness, resource efficiency and resource saving, as well as financial costs regarding the object of the study of the Master thesis. The competitiveness analysis is carried out for this purpose. The SWOT analysis helps to identify strengths, weaknesses, opportunities and threats associated with the project, and decide how to deal with them in each particular case. The development of the project requires funds that go to the salaries of project participants and the necessary equipment (the list is given in the respective section). The calculation of the resource efficiency indicator helps to make a final assessment of the technical decision on individual criteria and in general.

13.1 Competitiveness analysis of technical solutions

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

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

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

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

C - the competitiveness of research or a competitor;

Wi– criterion weight;

Pi – point of i-th criteria.

SIQ - smart interface quality;

EO - ease of operation;

ACPC - ability to connect to PC;

Table 5. Evaluation card for comparison of competitive technical solutions

Evaluation criteria ***example***	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_{SIQ}	P_{EO}	P_{ACPC}	C_{SIQ}	C_{EO}	C_{ACPC}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Energy efficiency	0.10	5	5	5	0.5	0.5	0.5
2. Reliability	0.20	4	4	5	0.8	0.8	1
3. Safety	0.10	5	5	5	0.5	0.5	0.5
4. Functional capacity	0.20	3	4	5	0.6	0.8	1
Economic criteria for performance evaluation							
1. Development cost	0.13	5	5	5	0.65	0.65	0.65
2. Market penetration rate	0.20	4	4	4	0.8	0.8	0.8
3. Expected lifecycle	0.07	4	4	5	0.28	0.28	0.35
Total	1	30	31	34	4.13	4.33	4.8

Make a conclusion according to the results of the competitiveness analysis.

13.2 SWOT analysis

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

Table 6. Matrix of SWOT-analysis

	Strengths: S1. Established a more detailed credit score tab S2. Provides detailed user ratings and an overview map of their default rates	**Weaknesses:** W1. There is a lot of noisy data in the data W2. Too many variables in the data set
Opportunities: O1. Has strong commercial value O2. Provide reference for the bank's credit scoring business	*Strategy which based on* *strengths and opportunities:* *1. The scorecard structure is simple and easy to implement*	*Strategy which based on* *weaknesses and opportunities:* *1. Use various data processing methods to deal with noisy data and improve the accuracy of the model*
Threats: T1. The scoring strategy derived from this project requires banks to implement differently according to their own business priorities	*Strategy which based on* *strengths and threats:* *1. Provides dual advice on maximizing profit and minimizing default rate*	*Strategy which based on* *weaknesses and threats:* *1. The credit score card obtained after cleaning the dirty data and the user's score and default rate profile have detailed information, which is convenient for the bank to formulate strategies*

13.3 Project Initiation

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

Table 7. Stakeholders of the project

Project stakeholders	Stakeholder expectations
Bank	The accuracy, reliability and availability of credit score cards
Bank customers	The specific behaviors and credit scoring framework that affect the credit scoring of bank customers

Table 8. Purpose and results of the project

Purpose of project:	Analyze the various behaviors of bank customers, establish a scoring card that can quantitatively display the credit status of bank customers, and provide reference for the bank's lending business.
Expected results of the project:	Established a more detailed credit score tab, provides detailed user ratings and an overview map of their default rates.
Criteria for acceptance of the project result:	Ensure that bank users who pass the bank loan review have a default rate of less than 10%
Requirements for the project result:	1. The project must be completed by June 1, 2021 of the year.
	2. The results obtained must meet the acceptance criteria for the project result.

The organizational structure of the project

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

Table 9. Participant of the project

№	Participant	Role in the project	Functions	Labor time, hours.
1	Supervisor	Head of project	Consultations. Review master's dissertation.	255 hours
2	Master's student	Executor	Writing master's dissertations. Through data mining, machine learning, credit score card related knowledge, gradually analyze bank customer data, and finally establish bank user credit score card Analyze and verify the accuracy of the credit score card.	765 hours

Project limitations

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

Table 10. Project limitations

Factors	**Limitations / Assumptions**
3.1. Project's budget	135000 RUB
3.1.1. Source of financing	TPU
3.2. Project timeline:	10/1/2021 to 24/05/2021
3.2.1. Date of approval of plan of project	20/03/2021
3.2.2. Completion date	24/05/2021

Project Schedule

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

Table 11. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
General Technical supervision	**30 days**	10/01/2021	8/02/2021	Supervisor
Research and analysis of literature	**30 days**	9/02/2021	10/03/2021	Supervisor/ Student
Clean data and build machine learning models	**30 days**	11/03/2021	9/04/2021	Supervisor/ Student
Build a credit score card based on the machine learning model	**30 days**	10/04/2021	09/05/2021	Supervisor/ Student
Preparing of dissertation	**15days**	10/05/2021	24/05/2020	Student

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

Table 12. Gantt chart

№	Activities	Participants	T_c, days	Duration of the project														
				January			February			March			April			May		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	General Technical supervision	Supervisor	30		▨	▨	▨											
2	Research and analysis of literature	Supervisor/ Student	30				▨	▨	▨	▨								
3	Clean data and build machine learning models	Supervisor/ Student	30								▨ ■	▨ ■	▨ ■					
4	Build a credit score card based on the machine learning model	Supervisor/ Student	30											▨ ■	▨ ■	▨ ■		
5	Preparing of dissertation	Student	15														▨ ■	▨ ■

13.4 Scientific and technical research budget

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

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

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

- material costs of scientific and technical research;
- costs of special equipment for scientific work (Depreciation of equipment used for design);
- basic salary;
- additional salary;
- labor tax;
- overhead.

13.5 Calculation xof material costs

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

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

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

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

P_i – the acquisition price of a unit of the i-th type of material resources consumed (rub./units, rub./kg, rub./m, rub./m^2, etc.);

k_T – coefficient taking into account transportation costs.

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

Energy costs are calculated by the formula:

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

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

P – power of equipment, kW;

F_{eq} – equipment usage time, hours.

Table 13. Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Electricity of computer	kWh	180	5.8	1044
Papers		120	1	120
Pen		2	150	300
Printing on A4 sheet		200	4	800
Internet	Month	6	350	2100
Total				4364

13.6 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:

$$S_b = S_a \cdot T_w, \qquad (3.3)$$

where S_b – basic salary per participant;

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

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

The average daily salary is calculated by the formula:

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

M – the number of months of work without leave during the year: at holiday in 48 days, M = 10.4 months, 6 day per week;

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.

Table 14. The valid annual fund of working time

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

Monthly salary is calculated by formula:

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

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 15. Calculation of the base salaries

Performers	S_{base}, rubles	$k_{premium}$	k_{bonus}	k_{reg}	S_{month}, rub.	W_d, rub.	T_p, work days	W_{base}, rub.
Head of project	35120			1.3	45656	1891.7	120	227004
Student	17310				22503	932.4	105	97902

13.7 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:

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

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient;

W_{base} – base salary, rubles.

13.8 Labor tax

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

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

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

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conducting educational and scientific activities have rate - 27.1%.

Table 16. Labor tax

	Project leader	**Engineer**
Coefficient of deductions	27.1%	
Salary, rubles	227004	97902
Labor tax, rubles	61518.1	26531.4

13.9 Overhead costs

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

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

Overhead is calculated according to the formula:

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

where k_{ov} – overhead rate.

Table 17. Overhead

	Project leader	**Engineer**
Overhead rate	30%	
Salary, rubles	227004	97902
Overhead, rubles	68101.2	29371

13.10 Formation of budget costs

The calculated cost of research is the basis for budgeting project costs.

Determining the budget for the scientific research is given in the table 15.

Table 18. Items expenses grouping

Name	**Cost, rubles**
1. Material costs	4364
2. Costs of special equipment	0
3. Basic salary	324906
4. Additional salary	0
5. Labor tax	88049.5

6. Overhead	97472.2
Total planned cost	514791.7

13.11Conclusion

Thus, in this section we developed stages for the design and creation of the competitive development that meets the requirements of the field of resource efficiency and resource saving.

These stages include:

- development of the economic project idea, formation of the project concept;
- organization of the work on the research project;
- identification of possible research alternatives;
- research planning;
- assessing the commercial potential and prospects of scientific research from the standpoint of resource efficiency and resource saving;
- determination of resource (resource saving), financial, budget, social and economic efficiency of the project.

Reference:

1. Brendan Jayagopal, N (2004). Applying Data Mining Techniques to Credit Scoring. Amadeus Software Limited.
2. Tatiana Inkhireeva, R (2016). Applying Data Mining Techniques to Credit Scoring.
3. Kaggle, [DB/OL] (2012). Give Me Some Credit: https://www.kag gle.com/c/GiveMeSomeCredit/overview
4. Ma Dongshenshen, [DB/OL] (2019). Credit card transform: https://zhuanlan.zhihu.com/p/90283372
5. Seeker, [DB/OL] (2018). Credit card understanding: https://zhuanlan.zhihu.com/p/36263276
6. August, [DB/OL] (2018). Playing with the financial scorecard model of logistic regression: https://zhuanlan.zhihu.com/p/36539125
7. Thomas LC, N (2000) A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International journal of forecasting 16(2), pp. 149–72
8. R.A. Eisenbeis, N (1978) Problems in applying discriminant analysis in credit scoring models. Journal of Banking and Finance, 2, pp. 205-219
9. Tina | Kinden Property, [DB/OL] (2019). Introduction of the Credit score card and WOE and IV: https://towardsdatascience.com/intro-to-credit-scorecard-9afeaaa3725f
10. Hui JiaowanUsing, [DB/OL] (2017). Calculation and use of IV value: https://www.jianshu.com/p/cc4724a373f8
11. JitendraShreemali N (2020), Comparing performance of multiple classifiers for regression and classification machine learning problems using structured datasets. Material Proceedings
12. Jiawei Yang N (2021), Mean-shift outlier detection and filtering. Pattern

Recognition 115 (2021)

13. Pablo Martínez Camblor N (2013), General nonparametric ROC curve comparison. Journal of the Korean Statistical Society 42 (1), pp. 71-81
14. GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements.
15. SP 2.4.3648-20. Sanitary and Epidemiological Requirements for Organizations of Education and Training, Recreation and Recreation of Children and Youth
16. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.
17. SanPiN 2.2.1 / 2.1.1.1278-03. Hygienic requirements for natural, artificial and combined lighting of residential and public buildings.
18. SanPiN 2.2.4.1329-03 Requirements for protection of personnel from the impact of impulse electromagnetic fields
19. GOST 12.1.038-82 Occupational safety standards system. Electrical safety
20. Federal Law "On the Fundamentals of Labor Protection in the Russian Federation" of 17.07.99 № 181 – FZ
21. GOST R ISO 1410-2010. Environmental management. Assessment of life Cycle. Principles and structure.
22. GOST R 52105-2003 Resources saving. Waste treatment.

List of abbreviations

WOE: Weight of Evidence

IV: Information Value

TPR: In real positive cases, the proportion of correct predictions

FPR: In real counter-examples, the percentage of correct predictions

ROC: Receiver Operating Characteristic)

AUC: Area Under the Curve

PDO: Point-to-Double Odds

Odd: Good to bad ratio

Offset: Offset value

Coef: Coefficient

List of tables

List of figures