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 Computer Science & Robotics
Field of training (Specialty): 09.04.04. Software engineering
Division: Information Technology

MASTER'S GRADUATION THESIS

Topic of research work
Analysis and prediction of behavioral characteristics of 4G users replacing 5G packages

UDC: 004.65:004.451:004.75:004.451

Student

Group	Full name	Signature	Date
8ПМ0И	Zhang Lifang		

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	Menshikova E.V.	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
Associate professor	Saveliev A.O.	PhD		

Томск – 2022

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 – 2022

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

School: School of Computer Science & Robotics
Field of training (Specialty): 09.04.04. Software engineering
Division: Information Technology

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

ASSIGNMENT
for the Graduation Thesis completion

In the form:

Master's Dissertation

For a student:

Group	**Full name**
8ПМ0И	Zhang Lifang

Topic of research work:

Analysis and prediction of behavioral characteristics of 4G users replacing 5G packages	
Approved by the order of the Director of School of Information Tech & Robotics (date, number):	25.05.2022, № 145-46/c

Deadline for completion of Master's Graduation Thesis:	06.10.2022

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.)*	In this work data mining helps to refers to the process of searching for information hidden in a large amount of data through algorithms. In this work data mining plays an extremely important role in useroriented Internet products

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. Find the right data source 2. Methods of analyzing data 3. Modeling with machine learning 4. Predicted results 5. Work on the section on financial management, resource efficiency and resource saving. 6. Work on the section on social responsibility.

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	Menshikova E.V
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
8ПМ0И	Zhang Lifang		

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 Computer Science & Robotics
Field of training (specialty): 09.04.04 «Software engineering»
Level of education: Master Degree Program
Division: Information Technology
Period of completion 2020/2021 and 2021/2022 academic years Form of presenting the work:
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:	15.06.2022

Assessment date	Title of section (module) / type of work (research)	Maximum score for the section (module)
27.01.2022	1. Preparation of technical specifications and selection ofresearch areas	
24.02.2022	2. Development of a common research methodology	
23.03.2022	3. Selection and study of materials on the topic	
13.04.2022	4. Obtaining necessary data and verification of the obtained results	
27.04.2022	5. Processing received data	
18.05.2022	6. Registration of the work performed	
26.05.2022	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	Saveliev A.O	Ph.D		

TASK FOR SECTION

«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING»

To the student:

Group	Full name
8ПМ0И	Zhang Lifang

School	Computer Science & Robotics	**Division**	Information Technology
Degree	Master	**Educational Program**	09.04.04 Software Engineering

Input data to the section «Financial management, resource efficiency and resource saving»:	
1. *Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human*	– Salary costs – 900000 – STR budget – 469074 – Other expenses - 0 rub.
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs – 5,8 rub per 1 kW
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Labor tax – 27,1 %; – Overhead costs – 16%; – District coefficient - 1.3
The list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	*Assessment of potential consumers of the study, SWOT - analysis,*
2. *Formation of a plan and budget for an engineering project*	*Planning the stages of work, determining the labor intensity and building a calendar schedule, budgeting.*
3. *Evaluation of resource, financial, social, budgetary effectiveness of IR and potential risks*	*Evaluation of the comparative effectiveness of the study.* *Integral indicator of resource efficiency - 4.1* *Integral efficiency indicator - 4.19* *Comparative efficiency of the project - 0.87*
A list of graphic material *(with list of mandatory blueprints):*	
1. *Competitiveness analysis* 2. *SWOT- analysis* 3. *Gantt chart and budget of scientific research* 4. *Assessment of resource, financial and economic efficiency of STR* 5. *Potential risks*	

Date of issue of the task for the section according to the schedule	30.05.2022

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	Menshikova E.V.	PhD		30.05.2022

The task was accepted by the student:

Group	Full name	Signature	Date
8ПМ0И	Zhang Lifang		30.05.2022

TASK FOR CHAPTER
«SOCIAL RESPONSIBILITY»

Student:

Group	Name		
8ПМ0И	Zhang Lifang		
School	Computer Science & Robotics	**Division**	Information Technology
Educational level	Master degree	**Course/Specialty**	09.04.04. Software Engineering

Topic of FQW:

Analysis and prediction of behavioral characteristics of 4G users replacing 5G packages	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	– Screen credit customers, according to the personal information on the finance dataset. – Model and evaluate accuracy using random forest and decision tree models.
List of questions to be researched, designed and developed:	
1. Legal and organizational issues of occupational safety – consider special (specific to the projected work area) law norms of labor legislation. – indicate the features of the labor legislation in relation to the specific conditions of the project.	– GOST 12.2.032-78 SSBT. Workplace when performing work while sitting General ergonomic requirements. – SanPiN 2.2.2/2.4.1340-03. Hygienic requirements for personal electronic computers and organization of work.
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the source of factor, the impact on human's body 2.2. Suggest measures to reduce the impact of identified harmful and dangerous factors	– Lack or lack of natural light, insufficient illumination. – Electromagnetic fields. – Excessive levels of noise. – Physical overload (static-long-term preservation of a certain posture).
3. Environmental Safety: Influence on the atmosphere, hydrosphere, lithosphere	– Hydrosphere: Computer components that contain hazardous materials: lead, cadmium, lithium, alkaline manganese, and mercury. – Lithosphere: Computer components that contain plastic, glass, lead, barium, and rare earth metals.
4. Emergency Safety: Describe the most likely emergency situation	– Fire

Date issue of the task for the chapter	

Consultant:

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

Student:

Group	**Name**	**Date**	**Signature**
8ПМ0И	Zhang Lifang		

Contents

1. Project Introduction 10

2. Data introduction 12

3. Prepare dataset 17

3.1. Check and preprocess the data 17

3.2 Data Classification 20

3.3 Split the data 20

4. User characteristics and behavior analysis 22

4.1 ANOVA 22

4.2 Categorical feature analysis 23

4.3 Numerical feature analysis 34

4.4 Summary of customer behavior characteristics 42

5. Model building and prediction 43

5.1 Create Transformers for Data Processing 43

5.2 Create a converter that selects columns 44

5.3 Building a data preprocessing pipeline 44

5.4 GradientBoostingClassifier 45

5.5 RandomForestClassifier 47

5.6 DecisionTreeClassifier 48

5.7 Model application 50

6. Conclusion 52

7. Financial management, resource efficiency and resource saving 53

8. Social responsibility 69

Reference 79

1. Project Introduction

The 5th Generation Mobile Communication Technology (5G for short) is a new generation of broadband mobile communication technology with high speed, low latency and large connection, and it is the network infrastructure that realizes the interconnection of human, machine and things.

The International Telecommunication Union (ITU) has defined three major application scenarios for 5G, namely enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (uRLLC), and massive machine type communication (mMTC)[5]. Enhanced Mobile Broadband (eMBB) is mainly for the explosive growth of mobile Internet traffic, providing mobile Internet users with a more extreme application experience; ultra-reliable and low-latency communication (uRLLC) is mainly for industrial control, telemedicine, and autonomous driving. Vertical industry application requirements with extremely high requirements for reliability and reliability; Massive Machine Type Communication (mMTC) is mainly for applications such as smart cities, smart homes, and environmental monitoring that target sensing and data collection.[1]

In recent years, social transformation has accelerated, and the country is strengthening the cultivation of the data factor market, promoting the modernization of the governance system, promoting the construction of new infrastructure, and striving to build a new smart city. The large-scale connection capabilities and high-speed transmission capabilities of 5G networks are the strong support for the construction of smart cities.

5G features high reliability, low latency, and large bandwidth, which can efficiently connect and integrate urban systems and services, improve resource utilization efficiency, optimize urban management and services, and improve the quality of life of citizens. Accelerating the in-depth integration of 5G user growth and urban development, and solving the problems brought about by the process of urbanization through informatization means is not only required for the sustainable

development of cities, but also where the new kinetic energy of the industry lies. How to accurately identify potential users of 5G demand through the model, promote the transition from 4G era to 5G era, and realize the construction of smart city based on 5G in-depth application is very important[2].

Based on the monthly user replacement 5G package data, analyze the behavioral characteristics of 4G users changing 5G packages, and construct 5G package potential customers from the dimensions of basic information, consumption behavior, super package information, broadband information, and other information of 4G users who replace 5G packages. Identify the model, identify the current 4G users who have the need to replace the 5G package, and carry out 5G potential customer marketing, as the vanguard of 5G smart city creation.

Due to the rapid development of 5g technology, we need to intensify efforts to promote 5g. We can analyze customer behavior characteristics and predict potential customers from existing customer data.

2. Data introduction

The data comes from China Mobile's big data platform, with a total data volume of more than 20W, including 44 columns of variable information. We can build a classification model, and desensitize the user number and user_id at the same time. The training set is 14W, the test set 6W.

The general data situation is shown in the following table.

Table 1. The general data situation

Dimension	Field Name	English meaning	Type of data	Remark
User	User ID	User ID	Brigint	User identification data will be desensitized
	Product_no	Product_no	Srting	
User basic information	X_1	Gender	Srting	
	X_2	Age	Int	
	X_3	Star	Int	
	X_4	Online time	Int	
	X_5	Market segment	Srting	
User consumption behavior	X_6	Current month arpu	Decimal(10,2)	Monthly discounted consumption information
	X_7	Last month arpu	Decimal(10,3)	Last month's discounted consumption information
	X_8	Last two-month arpu	Decimal(10,4)	Last two month's discounted

				consumption information
	X_9	Current month dou	Decimal(10,2)	Monthly traffic usage (unit: M)
	X_{10}	Last month dou	Decimal(10,2)	Traffic usage in the last month (unit: M)
	X_{11}	Last two month arpu	Decimal(10,2)	Traffic usage in the last two month (unit: M)
	X_{12}	Current month mou	Decimal(10,2)	Monthly voice usage (unit: minutes)
	X_{13}	Last month mou	Decimal(10,2)	Last month's voice usage (unit: minutes)
	X_{14}	Last two months mou	Decimal(10,2)	Last two month's voice usage (unit: minutes)
	X_{15}	The average arpu in the past three months	Decimal(10,2)	Average discounted consumption information in the past three months
	X_{16}	The average dou in the past three months	Decimal(10,2)	Average traffic usage information in the past three months

	X_{17}	The average mou in the past three months	Decimal(10,2)	Average minutes of voice calls used in the past three months
Over-set information	X_{18}	Voice over-set amount of the current month	Decimal(10,2)	
	X_{19}	Voice over-set amount of last month'	Decimal(10,2)	
	X_{20}	Voice over-set amount of last two month	Decimal(10,2)	
	X_{21}	Current months traffic over-set amount	Decimal(10,2)	
	X_{22}	Last month’s traffic over -set amount	Decimal(10,2)	
	X_{23}	Last two month’s traffic over-set amount	Decimal(10,2)	
Broadband Information	X_{24}	Whether the broadband user of this network	Smallint	
	X_{25}	Whether the broadband user of the different network	Smallint	

	X_{26}	Broadband bandwidth	Int	
	X_{27}	Whether the broadband is activated	Smallint	
Contract information	X_{28}	Broadband bundling contract ID	Smallint	
	X_{29}	Terminal bundle contract ID	Smallint	
	X_{30}	Call fee subscription ID	Smallint	
	X_{31}	Package subscription ID	Smallint	
Package information	X_{32}	User total package value (including theme packages and optional packages, etc.)	Decimal(10,2)	
	X_{33}	User main tariff package	Decimal(10,2)	
Flow Saturation Information	X_{34}	User traffic saturation of the current month	Decimal(10,2)	
	X_{35}	User traffic saturation of last month	Decimal(10,2)	

	X_{36}	User traffic saturation of last two month	Decimal(10,2)	
Other information	X_{37}	Whether it is a home user	Smallint	
	X_{38}	5G traffic	Smallint	
	X_{39}	terminal type	Smallint	
	X_{40}	whether the insurance account user is offset in the current month	Smallint	
	X_{41}	whether the phone is changed in the current month	Smallint	
	X_{42}	whether the place of residence covers the 5g logo	Smallint	
	X_{43}	whether the work place covers the 5g base station	Smallint	

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 [3]. If the machine learning model is used for analysis and prediction, it is necessary to divide the training set and Test set.

3.1. Check and preprocess the data

Code:

```
df_data.isnull().sum()
```

```
User ID                                                          0
product_no                                                       0
Gender                                                           0
Age                                                              0
Star                                                          6849
Online time                                                      0
Market segment                                                 691
current month arpu                                              89
last month arpu                                                 89
last two month arpu                                             89
current month dou                                               89
last month dou                                                  89
last two month dou                                              89
current month mou                                               89
last month mou                                                  89
last two months mou                                             89
The average arpu in the past three months                       89
the average dou in the past three months                        89
the average mou in the past three months                        89
Voice over-set amount of the current month                     392
Voice over-set amount of last month                            392
Voice over-set amount of last two month                        392
Current months traffic over-set amount                         392
Last months traffic over-set amount                            392
Last two months traffic over-set amount                        392
Whether the broadband user of this network                       0
Whether the broadband user of the different network              0
Broadband bandwidth                                         101060
Whether the broadband is activated                          101060
Broadband bundling contract ID                                 774
Terminal bundle contract ID                                    774
Call fee subscription ID                                       774
Package subscription ID                                        774
User total package value                                      6617
User main tariff package                                      6617
User traffic saturation of the current month                  9215
User traffic saturation of last month                         9793
User traffic saturation of last two month                     9585
Whether it is a home user                                        0
5G traffic                                                  132559
terminal type                                                    0
whether the insurance account user is offset in the current month  0
whether the phone is changed in the current month                0
whether the place of residence covers the 5g logo                0
whether the work place covers the 5g logo                        0
label                                                            0
dtype: int64
```

Figure 1. Check data

We need to check the degree of missing samples and the degree of missing features to determine whether to delete the missing samples.

Code:

```
threshold_features = 0.8

missing_col_num = missing_df[missing_df.missing_pct>=threshold_features].shape[0]

print ('The number of variables with a missing rate exceeding {} is {}'.format(threshold_features, missing_col_num))

plt.figure(figsize=(20, 5))

plt.title('Distribution map of missing features')

sns.barplot(data=missing_df[missing_df.missing_pct>0.01], x='col', y='missing_pct')

plt.ylabel('Missing rate')

plt.show()
```

```
The number of variables with a missing rate exceeding 0.8 is 1
```

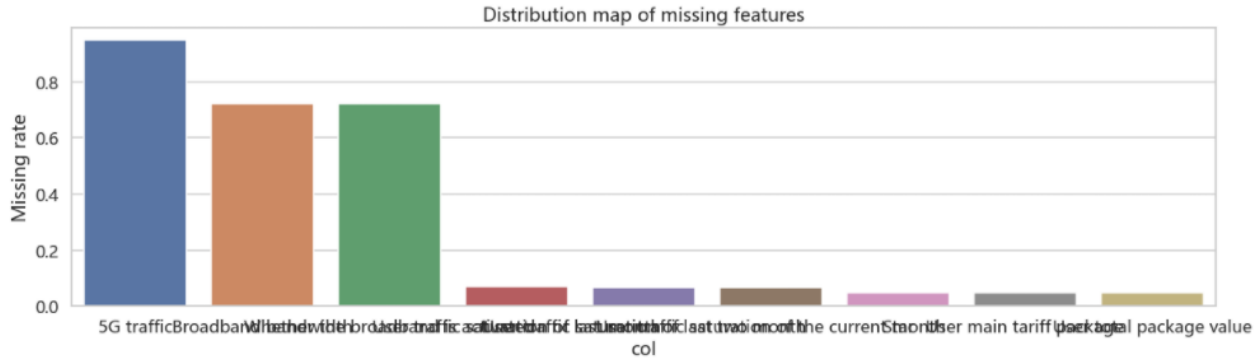


Figure 2. Missing degree of feature

Code:

```
missing_series = df_data.isnull().sum(axis=1)

list_missing_num  = sorted(list(missing_series.values))

plt.figure(figsize=(13, 5))

plt.title('Sample distribution plot of missing features')

plt.plot(range(df_data.shape[0]), list_missing_num)

plt.xlabel('samples')
```

```
plt.ylabel('number of missing features')

plt.show()
```

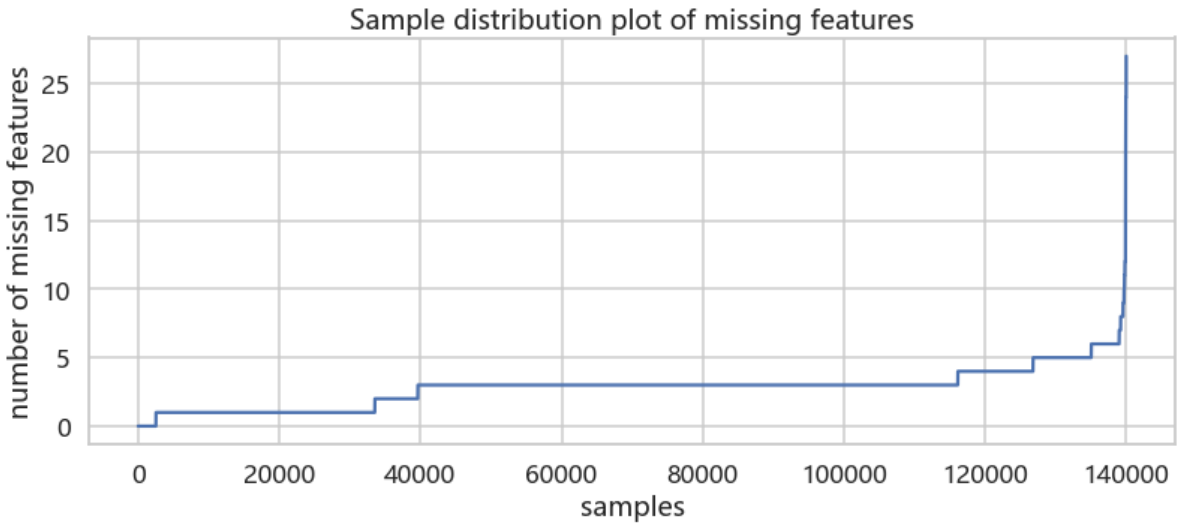


Figure 3. Missing degree of sample

There are 14w pieces of data in the training set, of which:

- User ID dimension has no missing data
- In the user basic information dimension, 6849 pieces of data are missing for star ratings, and 691 pieces of data for market segments are missing.
- All features in the dimension of consumption behavior information are missing 89 pieces of data, consider whether 89 users have missing consumption behavior
- All features in the dimension of super set information are missing 392 pieces of data, consider whether 392 users have super set information missing, and whether it is the same user with the missing consumer behavior
- There are 101,060 missing data in the broadband loan and broadband activation features in the broadband information dimension, and the missing is more serious
- All features in the contract information dimension are missing 774 data, consider the same super set of information dimensions

- All features in the package information dimension are missing 6617 pieces of data, considering the same super-set information dimension
- All features in the traffic saturation information dimension are missing 9000+ pieces of data, and further analysis is required

Among other information features, there are indeed 132,559 pieces of data in the 5G traffic feature, and the missing is very serious. Whether to delete this feature requires further analysis

Among the above-mentioned deletions, it is necessary to first consider whether the missing data of the same dimension belong to the same user; secondly, to further judge the degree of deletion by combining multiple dimensions; finally, to judge whether the feature can be directly eliminated for the feature with serious missing.

3.2 Data Classification

According to the above data preprocessing plan, the data of each column is classified.

Code:

```
num_columns=['X2', 'X3', 'X4', 'X15', 'X16', 'X17', 'X18_20', 'X21_23', 'X26', 'X32', 'X33', 'X34', 'X35', 'X36', 'X38'] #Numeric type

bool_columns=['X1', 'X24', 'X25', 'X27', 'X28', 'X29', 'X30', 'X31', 'X37', 'X40', 'X41', 'X42', 'X43'] # Boolean type

obj_columns=['X5', 'X39'] # Classification type
```

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.

In order to ensure the validity of the test data, here I use the stratified sampling method. Stratified sampling, also known as stratified extraction method, is a statistical method of sampling samples from a statistical population (also known as "matrix"). The sampling unit is divided into different layers according to a certain characteristic or a certain rule, and then samples are drawn independently and randomly from different layers. This ensures that the structure of the sample is relatively similar to the overall structure, thereby improving the accuracy of the estimation[4].

Code:

```
splt=StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
for train_idx, valid_idx in splt.split(data, label):
    train_part = data.iloc[train_idx,:]
    valid_part = data.iloc[valid_idx,:]
    train_part_label = label.iloc[train_idx]
    valid_part_label = label.iloc[valid_idx]
```

4. User characteristics and behavior analysis

4.1 ANOVA

ANOVA – In variance analysis, we call a certain characteristic of the object to be investigated as an experimental index, and the conditions that affect the experimental index are called factors. Factors can be divided into two categories. Factors such as professionalism); another type of factors that people cannot control (such as factors such as employee quality and opportunities). The factors discussed below are all controllable factors. Each factor has several states to choose from, and each state that the factor can choose from is called the level of the factor. If only one factor is changing in an experiment, it is called a univariate experiment; if more than one factor is changing, it is called a multivariate experiment[5].

Code:

```
threshold_const = 0.95

const_list = [x for x in df_data.columns if x!='label']

const_col = []

const_val = []

for col in const_list:

    max_samples_count = df_data[col].value_counts().iloc[0]

    sum_samples_count = df_data[df_data[col].notnull()].shape[0]


    const_val.append(max_samples_count/sum_samples_count)

    if max_samples_count/sum_samples_count >= threshold_const:

        const_col.append(col)


print('The number of features with constant variable/equivalence ratio greater than {} is {}'.format(threshold_const, len(const_col)))
```

```
plt.figure(figsize=(13, 5))
plt.title('Distribution map of isovalued features')
plt.plot(range(len(df_data.columns)-1), const_val)
plt.xlabel('number of features')
plt.ylabel('Equivalence ratio')
plt.show()
```

```
The number of features with constant variable/equivalence ratio greater than 0.95 is 5
```

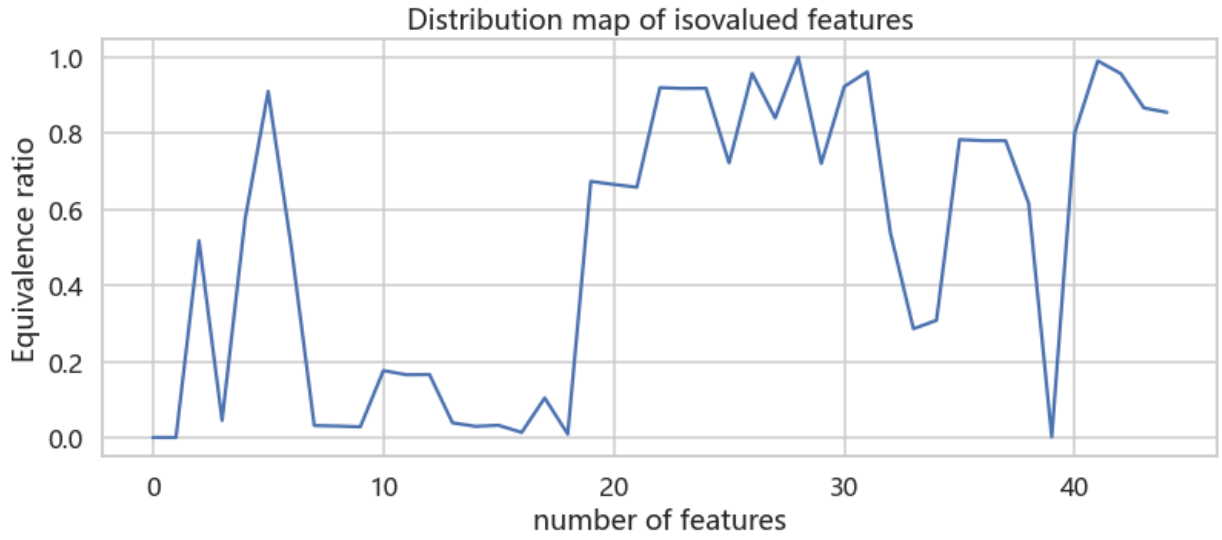


Figure 4. View the one-variance (isovalued) properties of features within features

Special attention should be paid to the 5 features with an equalization ratio greater than 0.95. There may be a feature that is all 1 value, or more than 99% is 1 value.

This feature is very helpful for model training and can be directly filtered out.

4.2 Categorical feature analysis

Categorical Feature – it mainly refers to features such as gender (male, female), blood type (A, B, AB, O) that only take values in limited options. The original input of categorical features is usually in the form of strings. Except for a

few models such as decision trees, the input of strings can be directly processed. For models such as logistic regression and support vector machines, categorical features must be processed and converted into numerical values to be correct. Work. When dealing with categorical features, it can be processed by encoding in various ways. Such as serial number encoding, one-hot encoding, binary encoding[6].

(1) Overview of Feature Analysis

Category features: gender, market segment, star rating, whether there is a broadband user on this network, whether a broadband user on a different network, broadband bandwidth, whether broadband is activated, broadband bundled contract sign, terminal bundle sign sign, phone bill sign sign, package sign sign,

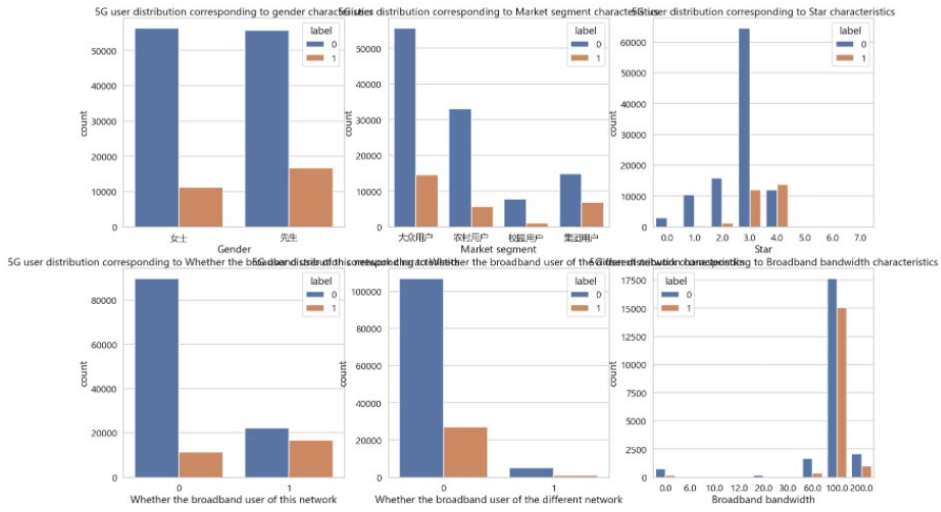
Whether it is a home user, the type of terminal, whether the account number is offset in the current month, whether the phone is changed in the current month, whether the place of residence covers the 5g mark, and whether the work place covers the 5g mark Analyze in turn.

Code:

```
fig, ax = plt.subplots(6, 3, figsize=(25, 50))

sns.countplot(x='Gender', hue='label', data=df_data, ax=ax[0, 0])

sns.countplot(x='Market segment', hue='label', data=df_data, ax=ax[0, 1])

sns.countplot(x='Star', hue='label', data=df_data, ax=ax[0, 2])

sns.countplot(x='Whether the broadband user of this network', hue='label', data=df_data, ax=ax[1, 0])

sns.countplot(x='Whether the broadband user of the different network', hue='label', data=df_data, ax=ax[1, 1])

sns.countplot(x='Broadband bandwidth', hue='label', data=df_data, ax=ax[1, 2])

sns.countplot(x='Whether the broadband is activated', hue='label', data=df_data, ax=ax[2, 0])
```

```
    sns.countplot(x='Broadband bundling contract ID', hue='label', data=df_data, ax=ax[2, 1])

    sns.countplot(x='Terminal bundle contract ID', hue='label', data=df_data, ax=ax[2, 2])

    sns.countplot(x='Call fee subscription ID', hue='label', data=df_data, ax=ax[3, 0])

    sns.countplot(x='Package subscription ID', hue='label', data=df_data, ax=ax[3, 1])

    sns.countplot(x='Whether it is a home user', hue='label', data=df_data, ax=ax[3, 2])

    sns.countplot(x='terminal type', hue='label', data=df_data, ax=ax[4, 0])

    sns.countplot(x='whether the insurance account user is offset in the current month', hue='label', data=df_data, ax=ax[4, 1])

    sns.countplot(x='whether the phone is changed in the current month', hue='label', data=df_data, ax=ax[4, 2])

    sns.countplot(x='whether the place of residence covers the 5g logo', hue='label', data=df_data, ax=ax[5, 0])

    sns.countplot(x='whether the work place covers the 5g logo', hue='label', data=df_data, ax=ax[5, 1])

    ax[0, 0].set_title('5G user distribution corresponding to gender characteristics')

    ax[0, 1].set_title('5G user distribution corresponding to Market segment characteristics')

    ax[0, 2].set_title('5G user distribution corresponding to Star characteristics')

    ax[1, 0].set_title('5G user distribution corresponding to Whether the broadband user of this network characteristics')

    ax[1, 1].set_title('5G user distribution corresponding to Whether the broadband user of the different network characteristics')

    ax[1, 2].set_title('5G user distribution corresponding to Broadband bandwidth characteristics')
```

```
    ax[2, 0].set_title('5G user distribution corresponding to Whether the broadband is activated characteristics')

    ax[2, 1].set_title('5G user distribution corresponding to Broadband bundling contract ID characteristics')

    ax[2, 2].set_title('5G user distribution corresponding to Terminal bundle contract ID characteristics')

    ax[3, 0].set_title('5G user distribution corresponding to Call fee subscription ID characteristics')

    ax[3, 1].set_title('5G user distribution corresponding to Package subscription ID characteristics')

    ax[3, 2].set_title('5G user distribution corresponding to Whether it is a home user characteristics')

    ax[4, 0].set_title('5G user distribution corresponding to terminal type characteristics')

    ax[4, 1].set_title('5G user distribution corresponding to whether the insurance account user is offset in the current month characteristics')

    ax[4, 2].set_title('5G user distribution corresponding to whether the phone is changed in the current month characteristics')

    ax[5, 0].set_title('5G user distribution corresponding to whether the place of residence covers the 5g logo')

    ax[5, 1].set_title('5G user distribution corresponding to whether the work place covers the 5g logo')

    plt.show()
```



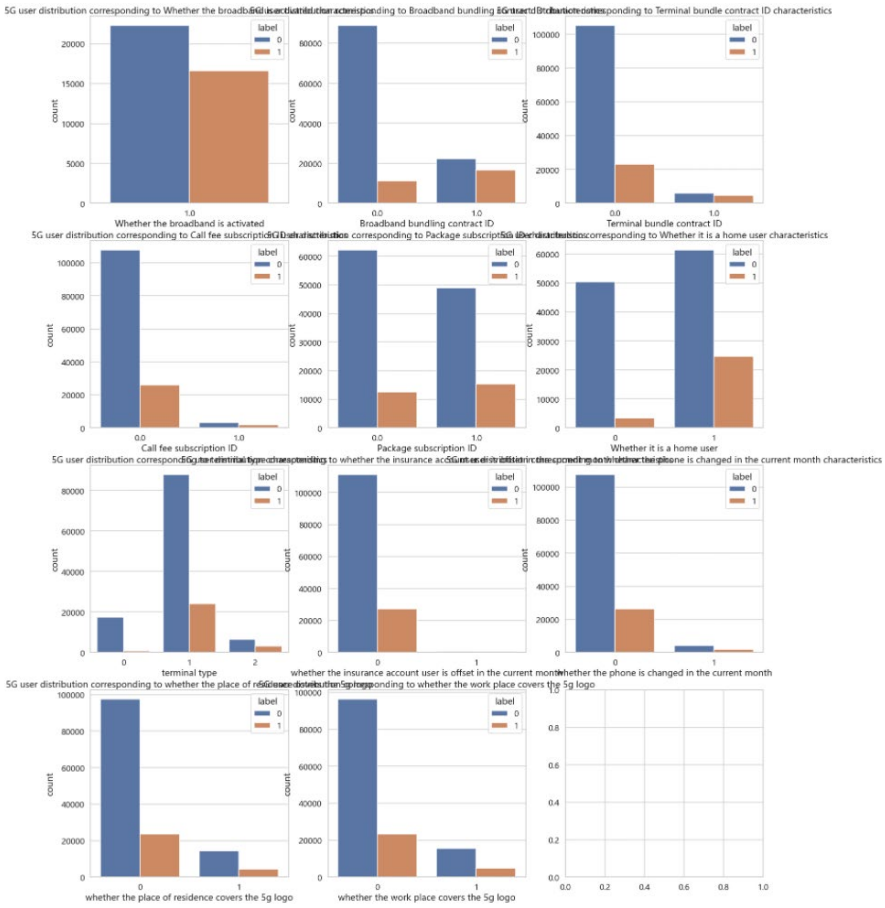


Figure 5. Category Characteristics Overview

From the above figure, we can draw the following conclusions:

- Among genders, male 5G user subscription rate is slightly higher than female
- In the market, group users are more inclined to turn on 5G
- Among the stars, four-star users are more willing to use 5G
- Broadband users on this network are more inclined to use 5G than broadband users on different networks, and the proportion of users with a broadband bandwidth of 100 is higher, followed by 200
- Users with broadband and terminal bundles are more willing to use 5G
- Home users are more willing to use 5G than non-home users
- Terminal type 1 has the most users, and terminal type 2 has the highest 5G user rate

in addition:

- Does broadband activate features with only unique and missing values
- Whether the account is maintained in the current month or not, the user characteristics are highly homogenized

The above features need to be further analyzed whether they can be deleted.

There are also: star rating and broadband bandwidth features account for a low proportion of individual features, which can be considered for binning; the above-mentioned features that tend to use 5G (for example: market segments) have null values in the categories. Is the missing value important? need further analysis

(2) Star feature analysis

Code:

```
df_bucket = df_data_2.groupby('Star')

user_5G_trend = pd.DataFrame()

user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

user_5G_trend
```

	Star	total	5G	5G_rate
0	0	2985	41	0.013735
1	1	10752	347	0.032273
2	2	17067	1194	0.069960
3	3	76459	11951	0.156306
4	4	25764	13755	0.533884
5	5	101	48	0.475248
6	6	19	9	0.473684
7	7	4	1	0.250000
8	empty	6849	654	0.095488

Figure 6. Star feature

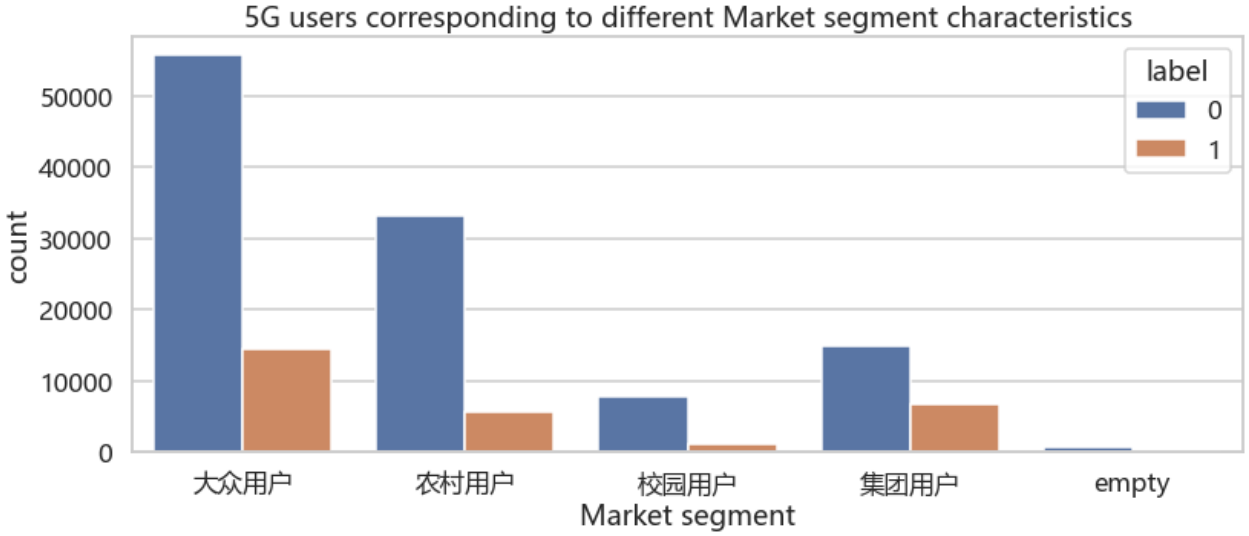
The proportion of 5G users corresponding to stars 0, 1, 2, 3, and 4 is gradually increasing, and the number of users of stars 5, 6, and 7 is relatively small, but the overall proportion of 5G users is still higher than the average box.

Correspondingly, the proportion of 5G users corresponding to the null value can be classified into one category separately, or it can be divided into bins and merged according to the proportion of 5G users, and the selection can be made according to the specific prediction effect of the model later.

(3) Market segment characteristics

Code:

```
df_data_2.loc[df_data_2['Market segment'].isnull(), 'Market segment'] = 'empty'

plt.figure(figsize=(13, 5))

plt.title('5G users corresponding to different Market segment characteristics')

sns.countplot(x='Market segment', hue='label', data=df_data_2)

plt.show()

df_bucket = df_data_2.groupby('Market segment')

user_5G_trend = pd.DataFrame()


user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

user_5G_trend
```



	Market segment	total	5G	5G_rate
0	empty	691	26	0.037627
1	农村用户	38742	5652	0.145888
2	大众用户	70076	14454	0.206262
3	校园用户	8786	1054	0.119964
4	集团用户	21705	6814	0.313937

Figure 7. Market characteristics and proportion

Campus users account for the least proportion and 5G activation rate is the lowest; group users have the highest 5G activation rate.

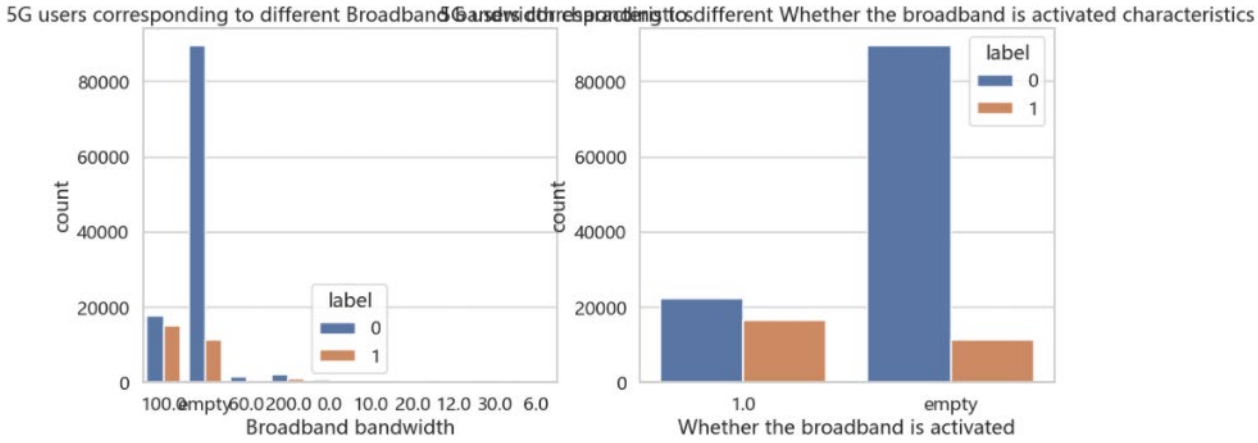
Correspondingly, the proportion of 5G users corresponding to the null value is extremely low, and the sub-box will affect the 5G activation rate of other boxes. Therefore, it is not recommended to combine the boxes. You can try to use it as a box alone. Later, it can be carried out according to the specific prediction effect of the model choose.

(4) Analysis of Broadband Bandwidth Characteristics

Code:

```
df_data_2.loc[df_data_2['Broadband bandwidth'].isnull(), 'Broadband bandwidth'] = 'empty'
```

```
df_data_2.loc[df_data_2['Whether the broadband is activated'].isnull(), 'Whether the broadband is activated'] = 'empty'

fig, ax = plt.subplots(1, 2, figsize=(15, 6))

sns.countplot(x='Broadband bandwidth', hue='label', data=df_data_2, ax=ax[0])

sns.countplot(x='Whether the broadband is activated', hue='label', data=df_data_2, ax=ax[1])

ax[0].set_title('5G users corresponding to different Broadband bandwidth characteristics')

ax[1].set_title('5G users corresponding to different Whether the broadband is activated characteristics')

plt.show()

df_bucket = df_data_2.groupby('Broadband bandwidth')

user_5G_trend = pd.DataFrame()


user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

user_5G_trend
```



	Broadband bandwidth	total	5G	5G_rate
0	0	936	179	0.191239
1	6	14	6	0.428571
2	10	23	6	0.260870
3	12	5	2	0.400000
4	20	184	38	0.206522
5	30	1	1	1.000000
6	60	1989	362	0.182001
7	100	32708	15053	0.460224
8	200	3080	987	0.320455
9	empty	101060	11366	0.112468

	Whether the broadband is activated	total	5G	5G_rate
0	1	38940	16634	0.427170
1	empty	101060	11366	0.112468

Figure 7. Activation of broadband bandwidth 5g users and their proportion

Broadband Bandwidth Features:

- Features with bandwidth below 60 can be binned because there is too little sample data
- If the bandwidth is above 100, because the overall proportion of 5G users is relatively high, it can be combined.
- That is, the final broadband bandwidth feature can be divided into 3 boxes: <=60, >100, empty

Whether broadband is activated or not:

Because the attribute is only 1, it means that the broadband has been activated, so boldly predicting the missing data here means that the broadband is not activated, and the corresponding can be represented by 0.

(5) Terminal bundling subscription identification and broadband bundling subscription identification features

Code:

```
df_data_2.loc[df_data_2['Broadband bundling contract ID'].isnull(), 'Broadband bundling contract ID'] = 'empty'

df_bucket = df_data_2.groupby('Broadband bundling contract ID')

user_5G_trend = pd.DataFrame()

user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

user_5G_trend

df_data_2.loc[df_data_2['Terminal bundle contract ID'].isnull(), 'Terminal bundle contract ID'] = 'empty'


df_bucket = df_data_2.groupby('Terminal bundle contract ID')

user_5G_trend = pd.DataFrame()

user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

user_5G_trend
```

	Broadband bundling contract ID	total	5G	5G_rate
0	0	100303	11304	0.112699
1	1	38923	16664	0.428127
2	empty	774	32	0.041344

	Terminal bundle contract ID	total	5G	5G_rate
0	0	128467	23182	0.180451
1	1	10759	4786	0.444837
2	empty	774	32	0.041344

Figure 8. Terminal bundling subscription identification and broadband bundling their proportion

Analysis of any two of the first four features shows that the proportion of 5G users with missing data is relatively low.

Provide the following two ideas:

- Merge into bins with similar probabilities, such as the 0 attribute in the above feature
- Fill with mode
- Fill with the mode of the feature of users in the same category
- Delete the sample directly

4.3 Numerical feature analysis

(1) Analysis of age characteristics

Calculate the distribution of 5G users by age.

Code:

```
df_bucket = df_data_2.groupby('Age')

user_5G_trend = pd.DataFrame()

user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()

user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

fig, ax = plt.subplots(2, 1, figsize=(25,10))

plt.title('Trend chart of the proportion of 5G users in different age groups')

sns.countplot(x='Age', hue='label', data=df_data.sort_values(['Age']), ax=ax[0])

sns.pointplot(data=user_5G_trend, x='Age', y='5G_rate', ax=ax[1])

plt.show()
```

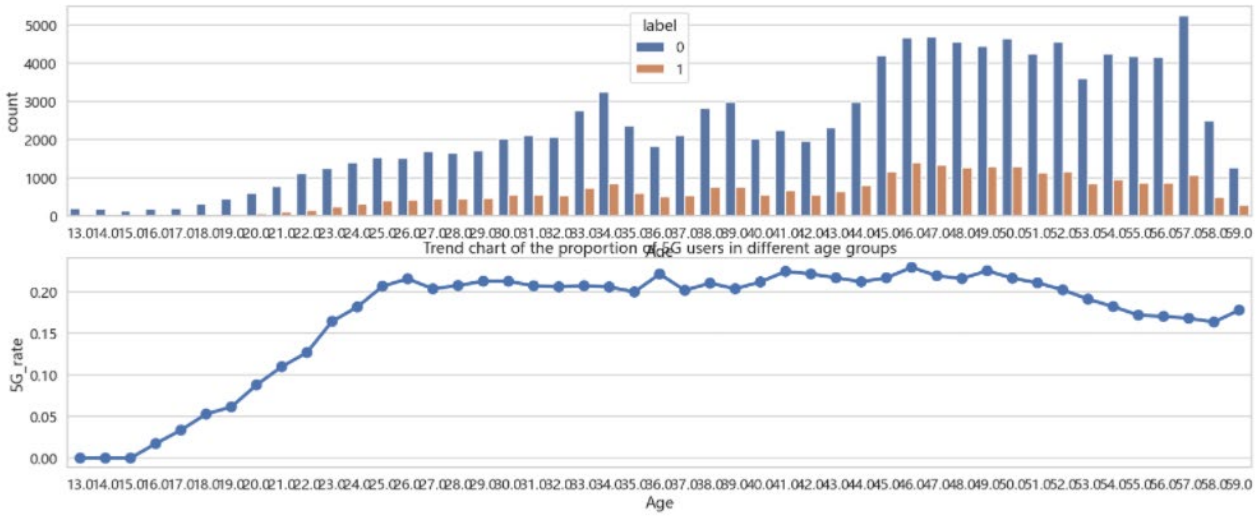


Figure 9. Age characteristics

Although age features are numerical, they can also be analyzed analogously to categorical features.

It can be found:

- The overall proportion of 5G users between the ages of 25 and 50 is relatively high, basically fluctuating above 20%
- The proportion of 5G users aged 13-25 is increasing with age
- The proportion of 5G users over 50 years old is declining with age

Based on this feature, age features can be binned, roughly as follows:

- ①13-20, ②20-25 ③25-45 ④ 45-50 ⑤>50

(2) Analysis of the characteristics of online time

Calculate the distribution of 5G users with different network durations.

Code:

```
df_bucket = df_data_2.groupby('Online time')

user_5G_trend = pd.DataFrame()

user_5G_trend['total'] = df_bucket['label'].count()

user_5G_trend['5G'] = df_bucket['label'].sum()
```

```
user_5G_trend['5G_rate'] = user_5G_trend['5G']/user_5G_trend['total']

user_5G_trend = user_5G_trend.reset_index()

fig, ax = plt.subplots(2, 1, figsize=(25,15))

plt.title('Trend chart of the proportion of 5G users with different online time')

sns.countplot(x='Online time', hue='label', data=df_data.sort_values(['Online time']), ax=ax[0])

sns.pointplot(data=user_5G_trend, x='Online time', y='5G_rate', ax=ax[1])

plt.show()
```

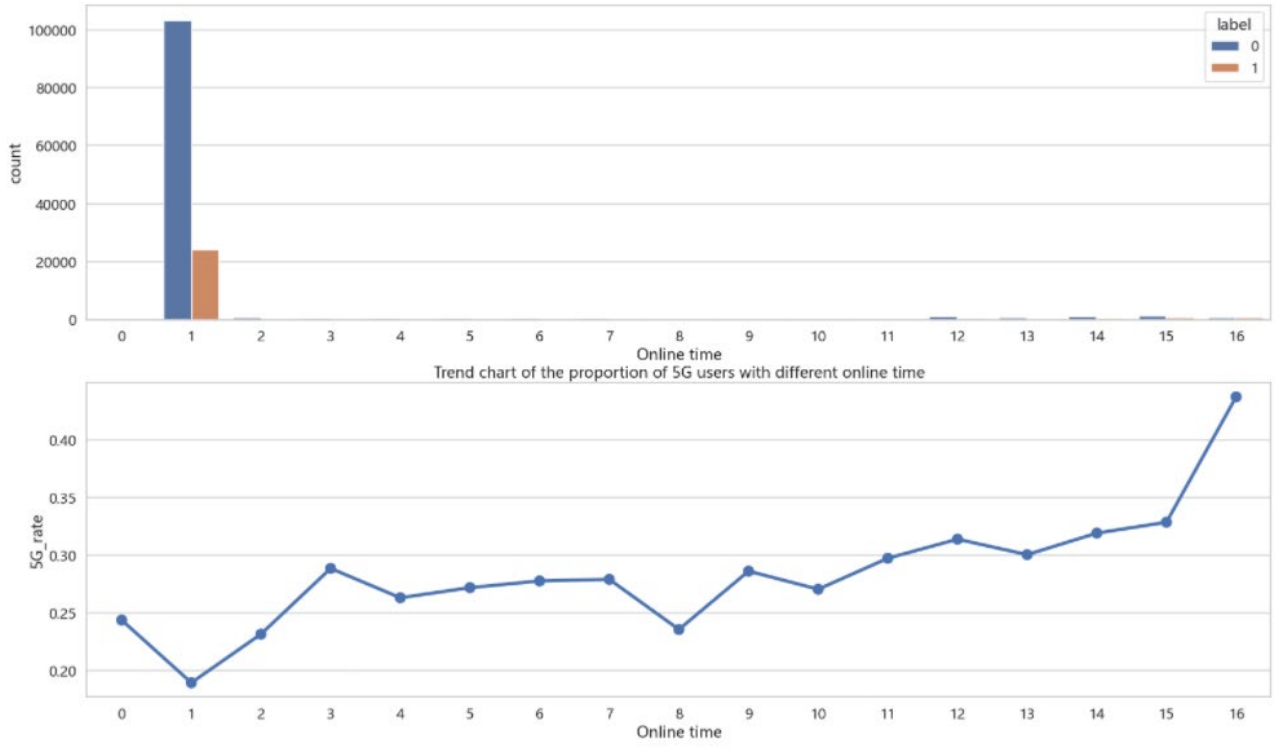


Figure 10. Online time characteristics

The sample with a network duration of 1 accounted for the largest proportion, but at the same time the proportion of 5G users was also the lowest;

When the online time is between 2-10, the proportion of 5G users fluctuates between 25%-30%;

When the online time is more than 10, the proportion of 5G users is higher than 30%, and it shows an accelerating upward trend.

In fact, after analyzing this, it is not difficult to guess that the corresponding number should represent the user's online year:

Specific operation suggestions: Or sub-box operation ①0-1 ② 2-10 ③>10

(3) Draw combined graphics(Box Plot)

It is also called discretization of continuous variables.

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 con tinuity of the variable determine the type of binning of the variable.

In this project, continuous variables can be optimized segmentation, and discontinuous variables can be manually binned.

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

Code:

```
fig, ax = plt.subplots(4, 3, figsize=(25, 28))

sns.boxenplot(x=np.ones(df_data.shape[0]), y='current month arpu', data=df_data, ax=ax[0, 0])

sns.boxenplot(x=np.ones(df_data.shape[0]), y='last month arpu', data=df_data, ax=ax[1, 0])

sns.boxenplot(x=np.ones(df_data.shape[0]), y='last two month arpu', data=df_data, ax=ax[2, 0])

sns.boxenplot(x=np.ones(df_data.shape[0]), y='current month dou', data=df_data, ax=ax[0, 1])

sns.boxenplot(x=np.ones(df_data.shape[0]), y='last month dou', data=df_data, ax=ax[1, 1])
```

```
    sns.boxenplot(x=np.ones(df_data.shape[0]), y='last two month dou', data=df_data, ax=ax[2, 1])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='current month mou', data=df_data, ax=ax[0, 2])


    sns.boxenplot(x=np.ones(df_data.shape[0]), y='last month mou', data=df_data, ax=ax[1, 2])


    sns.boxenplot(x=np.ones(df_data.shape[0]), y='last two months mou', data=df_data, ax=ax[2, 2])


    sns.boxenplot(x=np.ones(df_data.shape[0]), y='The average arpu in the past three months', data=df_data, ax=ax[3, 0])


    sns.boxenplot(x=np.ones(df_data.shape[0]), y='the average dou in the past three months', data=df_data, ax=ax[3, 1])


    sns.boxenplot(x=np.ones(df_data.shape[0]), y='the average mou in the past three months', data=df_data, ax=ax[3, 2])


    ax[0, 0].set_title('current month arpu distributed')

    ax[1, 0].set_title('last month arpu distributed')

    ax[2, 0].set_title('last two month arpu distributed')

    ax[0, 1].set_title('current month dou distributed')

    ax[1, 1].set_title('last month dou distributed')

    ax[2, 1].set_title('last two month dou distributed')

    ax[0, 2].set_title('current month mou distributed')

    ax[1, 2].set_title('last month mou distributed')

    ax[2, 2].set_title('last two month mou distributed')

    ax[3, 0].set_title('The average arpu in the past three months distributed')
```

```
    ax[3, 1].set_title('The average dou in the past three months distributed')

    ax[3, 2].set_title('The average mou in the past three months distributed')

    plt.show()
```

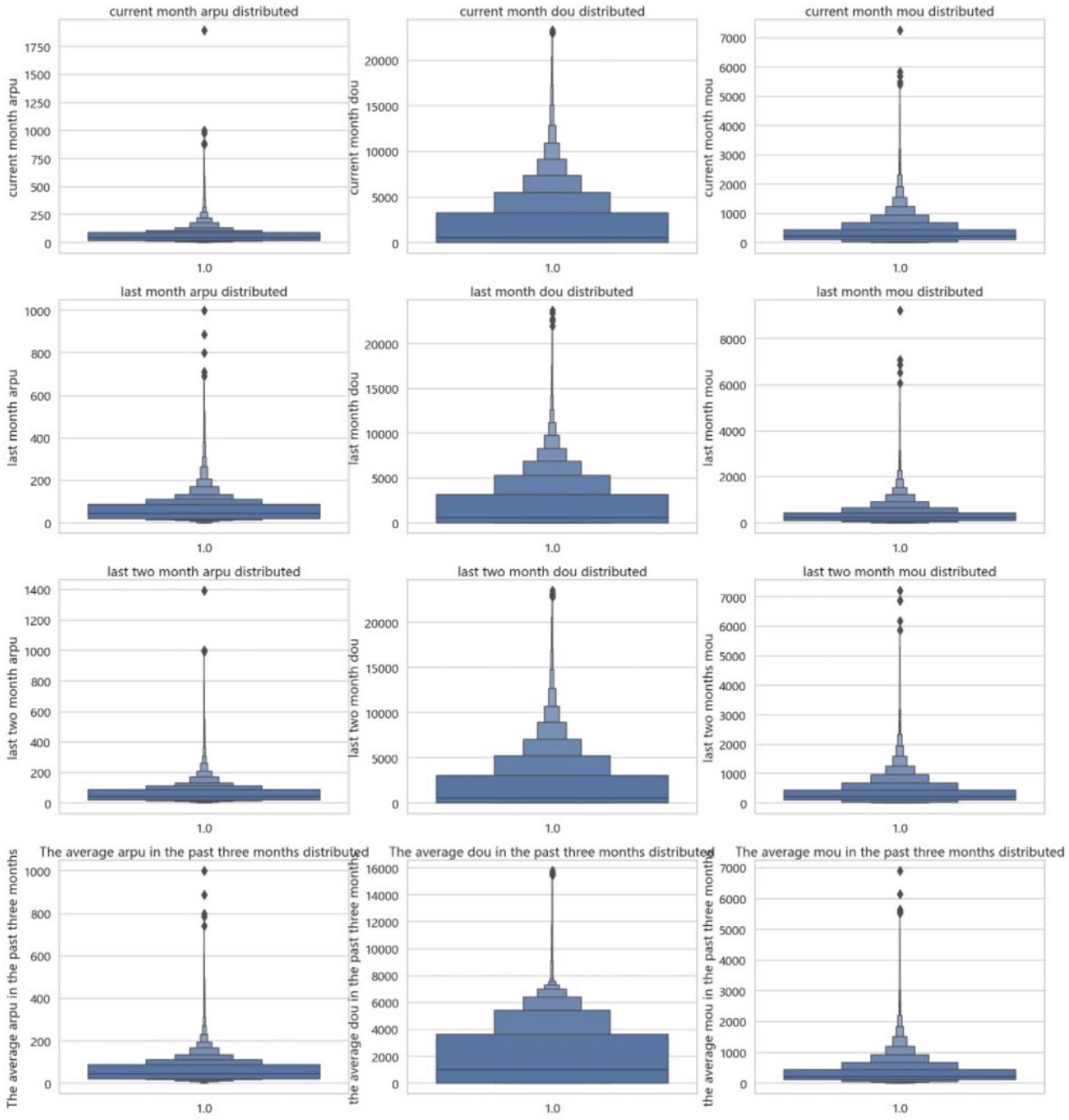


Figure 11. General numeric data binning

Code:

```
    fig, ax = plt.subplots(4, 3, figsize=(25, 28))
```

```
    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Voice over-set amount
of the current month', data=df_data, ax=ax[0, 0])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Voice over-set amount
of last month', data=df_data, ax=ax[1, 0])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Voice over-set amount
of last two month', data=df_data, ax=ax[2, 0])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='User traffic saturation
of the current month', data=df_data, ax=ax[0, 1])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='User traffic saturation
of last month', data=df_data, ax=ax[1, 1])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='User traffic saturation
of last two month', data=df_data, ax=ax[2, 1])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Current months traffic
over-set amount', data=df_data, ax=ax[0, 2])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Last months traffic
over-set amount', data=df_data, ax=ax[1, 2])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='Last two months traffic
over-set amount', data=df_data, ax=ax[2, 2])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='User total package
value', data=df_data, ax=ax[3, 0])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='User main tariff
package', data=df_data, ax=ax[3, 1])

    sns.boxenplot(x=np.ones(df_data.shape[0]), y='5G traffic',
data=df_data, ax=ax[3, 2])


    ax[0, 0].set_title('Voice over-set amount of the current month
distributed')

    ax[1, 0].set_title('Voice over-set amount of last month distributed')

    ax[2, 0].set_title('Voice over-set amount of last two month
distributed')

    ax[0, 1].set_title('User traffic saturation of the current month
distributed')

    ax[1, 1].set_title('User traffic saturation of last month distributed')
```

```
ax[2, 1].set_title('User traffic saturation of last two month distributed')

ax[0, 2].set_title('Current months traffic over-set amount distributed')

ax[1, 2].set_title('Last months traffic over-set amount distributed')

ax[2, 2].set_title('Last two months traffic over-set amount distributed')

ax[3, 0].set_title('User total package value distributed')

ax[3, 1].set_title('User main tariff package distributed')

ax[3, 2].set_title('5G traffic distributed')

plt.show()
```

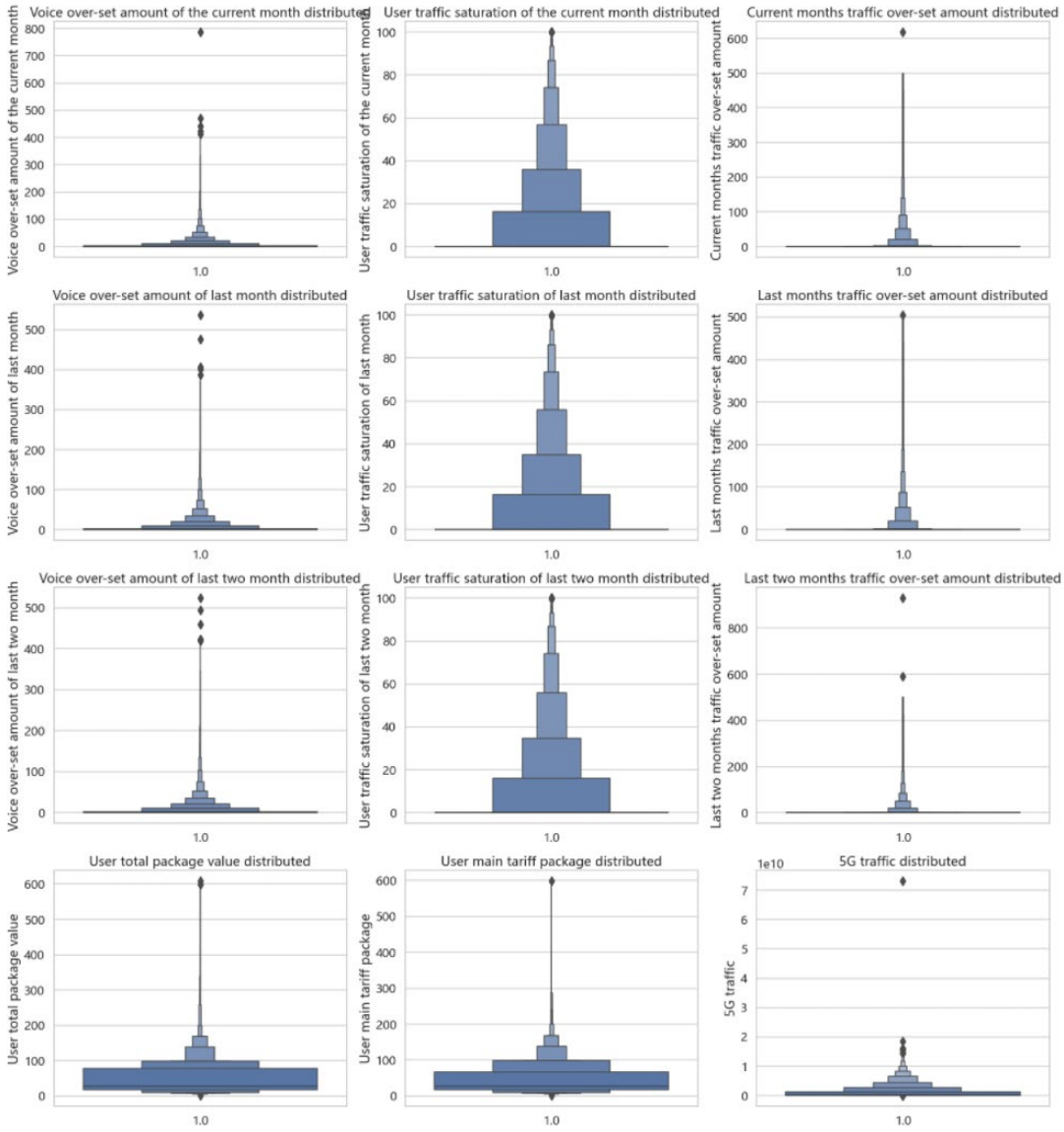


Figure 12. Other numeric data binning

4.4 Summary of customer behavior characteristics

The above is what needs to be done in the data exploration stage. Common exploration methods will also conduct simultaneous analysis of multiple features.

First it is obvious that:

- Among genders, male 5G user subscription rate is slightly higher than female
- In the market, group users are more inclined to turn on 5G
- Among the stars, four-star users are more willing to use 5G
- Broadband users of this network are more inclined to use 5G, and users with a broadband bandwidth of 100 are more willing
- Users with broadband and terminal bundles are more willing to use 5G
- Home users are more willing to use 5G than non-home users
- The 5G user rate with terminal type 2 is the highest

In a little more detail:

- The proportion of 5G users corresponding to stars 0, 1, 2, 3, and 4 gradually increases, and the number of users of stars 5, 6, and 7 is the least, but the proportion of 5G users is similar
- Campus users account for the least, and the 5G activation rate is the lowest; group users have the highest 5G activation rate
- The sample data with bandwidth below 60 is too small, and the proportion of 5G users with bandwidth above 100 is relatively high
- The property of whether broadband is activated is only 1, indicating that broadband has been activated.
- The proportion of 5G users with empty contract information in the contract dimension feature is relatively low

5. Model building and prediction

5.1 Create Transformers for Data Processing

According to the analysis of the fourth step, the following transformation processing is performed on some data

- For the index data, it is observed that user_id corresponds to product_no one-to-one, and user_id is used in the label data, so user_id is selected as the data index, and the product_no field is deprecated
- For arpu, dou, and mou, the average data of the first three months and the first three months are given at the same time, so the corresponding March average value is directly selected, and the data of each month is discarded.
- For voice over-set and traffic over-set, only the monthly data of the first three months is given. In order to unify with arpu, dou, and mou, the average value of three months is calculated for modeling, and the original data is discarded.
- For gender, Mr. and Ms. are used in the data. For the convenience of analysis and calculation, they are converted to 1 and 0 respectively.
- For segment fields, fill missing values with "Other"
- Other missing values in the data can be uniformly filled with 0
- For numerical data, normalization is required to improve data fitting efficiency
- For categorical data, OneHot transformation is required to facilitate subsequent fitting

Code:

```
class Preprocessing(BaseEstimator, TransformerMixin):
    def __init__(self, testset=False):
        self.testset = testset
        super().__init__()
    def fit(self, data, y=None):
        return self
```

```
    def transform(self, data):
        X=data.copy()
        X.set_index('user_id',inplace=True)
        X.loc[X['X1']=='先生','X1']=1
        X.loc[X['X1']=='女士','X1']=0
        X['X18_20']=(X['X18']+X['X19']+X['X20'])/3
        X['X21_23']=(X['X21']+X['X22']+X['X23'])/3
        X.drop(['product_no','X6', 'X7', 'X8', 'X9', 'X10', 'X11', 'X12',
'X13', 'X14', 'X18', 'X19', 'X20', 'X21', 'X22', 'X23'],axis=1,inplace=True)
        X.X5.fillna('other',inplace=True)
        X.fillna(0,inplace=True)
        return X
```

5.2 Create a converter that selects columns

Code:

```
class ColSelector(BaseEstimator, TransformerMixin):
    def __init__(self, attribute_names):
        self.attribute_names = attribute_names
    def fit(self, X, y=None):
        return self
    def transform(self, X):
        return X[self.attribute_names].values
```

5.3 Building a data preprocessing pipeline

Select numerical data and perform normalization, select categorical data, perform OneHot transformation, merge Boolean data, numerical data, categorical data, socket data preprocessing pipeline, and integrate into a complete processing pipeline.

Code:

```
pipe=Pipeline([
        ('prep', pipe_full),
        ('model', ())
    ])
```

In order to facilitate the selection of different algorithms, the preprocessing and fitting algorithms are separated when building the overall model, so that the selection of different algorithms can be realized.

5.4 GradientBoostingClassifier

Code:

```
pipe=Pipeline([
        ('prep', pipe_full),
        ('model', GradientBoostingClassifier())
    ])
params={
        'model__learning_rate' : [0.01, 0.1],
        'model__n_estimators' : [100,500],
        'model__max_features' : ['sqrt'],
        'model__max_depth' : [6,7,8],
        'model__min_samples_split' : [1000,1500],
        'model__min_samples_leaf' : [10,50,100],
        'model__subsample' : [1]
    }
splt = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
    for train_idx, valid_idx in splt.split(data, label):
        train_part = data.iloc[train_idx,:]
        valid_part = data.iloc[valid_idx,:]
```

```
        train_part_label = label.iloc[train_idx]

        valid_part_label = label.iloc[valid_idx]

grid_search = GridSearchCV(pipe, params, cv=3, scoring='roc_auc', verbose=2,
n_jobs=-1)
```

```
GridSearchCV(cv=3,
             estimator=Pipeline(steps=[('prep',
                                        Pipeline(steps=[('preprocessing',
                                                         Preprocessing()),
                                                        ('featureunion',
                                                         FeatureUnion(transformer_list=[('colselector',
                                                                                         ColSelector(attribute_names=['X1',
                                                                                                                      'X24',
                                                                                                                      'X25',
                                                                                                                      'X27',
                                                                                                                      'X28',
                                                                                                                      'X29',
                                                                                                                      'X30',
                                                                                                                      'X31',
                                                                                                                      'X37',
                                                                                                                      'X40',
                                                                                                                      'X41',
                                                                                                                      'X42',
                                                                                                                      'X43'])),
                                                                                        ('pipeline-1',
                                                                                         Pipeline(steps=[('colselector',
                                                                                                          ColSelector(attribute_names=['X
2...

                                                                                                          OneHotEncoder(sparse=Fals
e))])])])])),
                                        ('model',
                                         GradientBoostingClassifier())]),
             n_jobs=-1,
             param_grid={'model__learning_rate': [0.01, 0.1],
                         'model__max_depth': [6, 7, 8],
                         'model__max_features': ['sqrt'],
                         'model__min_samples_leaf': [10, 50, 100],
                         'model__min_samples_split': [1000, 1500],
                         'model__n_estimators': [100, 500],
                         'model__subsample': [1]},
             scoring='roc_auc', verbose=2)
```

Figure 13. Best parameters and models

Finally, we need to evaluate the model and use the test set to evaluate the accuracy of the model

Code:

```
final_model = grid_search.best_estimator_

score=roc_auc_score(valid_part_label,final_model.predict_proba(valid_part)
[:,1])
```

The final model achieved 90.87% accuracy on the test set, and it seems that the fitting results are good.

5.5 RandomForestClassifier

Code:

```
pipe=Pipeline([
        ('prep', pipe_full),
        ('model', RandomForestClassifier())
    ])
params={
        'model__n_estimators' : [100,500],
        'model__max_features' : ['sqrt'],
        'model__max_depth' : [6,7,8],
        'model__min_samples_split' : [1000,1500],
        'model__min_samples_leaf' : [10,50,100],
    }
splt = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
    for train_idx, valid_idx in splt.split(data, label):
        train_part = data.iloc[train_idx,:]
        valid_part = data.iloc[valid_idx,:]
        train_part_label = label.iloc[train_idx]
        valid_part_label = label.iloc[valid_idx]
grid_search = GridSearchCV(pipe, params, cv=3, scoring='roc_auc', verbose=2,
n_jobs=-1)
```

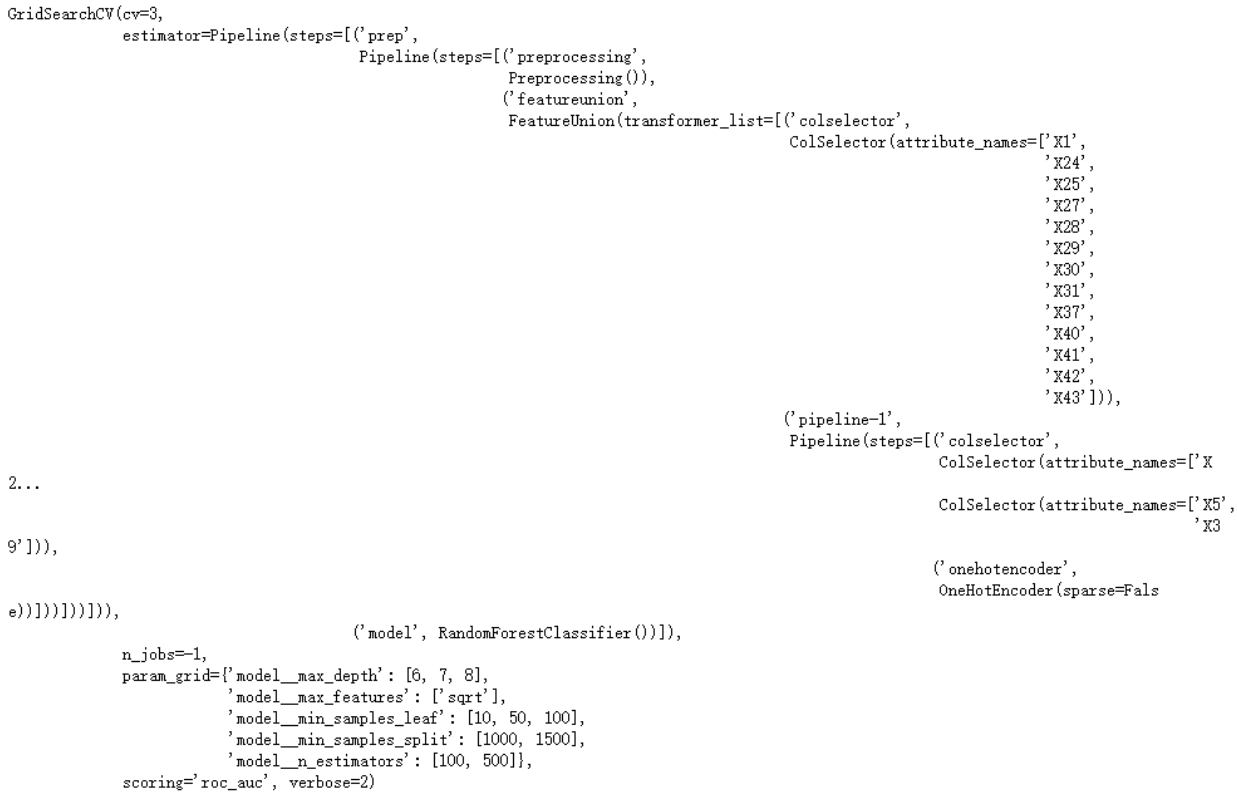


Figure 14. Best parameters and models

Finally, we need to evaluate the model and use the test set to evaluate the accuracy of the model

Code:

```
final_model = grid_search.best_estimator_

score=roc_auc_score(valid_part_label,final_model.predict_proba(valid_part)
[:,1])
```

The final model achieved 70.92% accuracy on the test set, and it seems that the fitting results are worse.

5.6 DecisionTreeClassifier

Code:

```
model_DTC = tree.DecisionTreeClassifier()

pipe=Pipeline([

        ('prep', pipe_full),
```

```
        ('model', model_DTC)
    ])
params={
        'model__max_depth' : [6,7,8],
        'model__min_samples_split' : [1000,1500],
        'model__min_samples_leaf' : [10,50,100],
    }
splt = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
    for train_idx, valid_idx in splt.split(data, label):
        train_part = data.iloc[train_idx,:]
        valid_part = data.iloc[valid_idx,:]
        train_part_label = label.iloc[train_idx]
        valid_part_label = label.iloc[valid_idx]
grid_search = GridSearchCV(pipe, params, cv=3, scoring='roc_auc', verbose=2,
n_jobs=-1)
```

```
GridSearchCV(cv=3,
             estimator=Pipeline(steps=[('prep',
                                        Pipeline(steps=[('preprocessing',
                                                         Preprocessing()),
                                                        ('featureunion',
                                                         FeatureUnion(transformer_list=[('colselector',
                                                                                         ColSelector(attribute_names=['X1',
                                                                                                                      'X24',
                                                                                                                      'X25',
                                                                                                                      'X27',
                                                                                                                      'X28',
                                                                                                                      'X29',
                                                                                                                      'X30',
                                                                                                                      'X31',
                                                                                                                      'X37',
                                                                                                                      'X40',
                                                                                                                      'X41',
                                                                                                                      'X42',
                                                                                                                      'X43'])),
                                                                                        ('pipeline-1',
                                                                                         Pipeline(steps=[('colselector',
                                                                                                          ColSelector(attribute_names=['X
2...
                                                                                                          StandardScaler())])),
                                                                                        ('pipeline-2',
                                                                                         Pipeline(steps=[('colselector',
                                                                                                          ColSelector(attribute_names=['X5',
                                                                                                                                       'X3
9'])),
                                                                                                         ('onehotencoder',
                                                                                                          OneHotEncoder(sparse=Fals
e))])])])])),
                                        ('model', DecisionTreeClassifier())]),
             n_jobs=-1,
             param_grid={'model__max_depth': [6, 7, 8],
                         'model__min_samples_leaf': [10, 50, 100],
                         'model__min_samples_split': [1000, 1500]},
             scoring='roc_auc', verbose=2)
```

Figure 15. Best parameters and models

Finally, we need to evaluate the model and use the test set to evaluate the accuracy of the model

Code:

```
final_model = grid_search.best_estimator_

score=roc_auc_score(valid_part_label,final_model.predict_proba(valid_part)[:,1])
```

The final model achieved 55.6% accuracy on the test set, and it seems that the fitting results are the worst.

5.7 Model application

Comparing the accuracy, it can be seen that GradientBoostingClassifier has the highest accuracy.

After getting the best model we need to practice with modeling data. Predict how many users are willing to upgrade from 4g to 5g plans. For this I have prepared the dataset `result_predict.csv`

Code:

```
test =pd.read_csv('result_predict.csv')

result =pd.DataFrame({'proba':final_model.predict_proba(test)[:,1]},index=test.user_id)

5G=result.loc[result.proba>0.5].count()

total=result.count()

rate= G/ total

print (' There are %d pieces of data in test set. According to the model prediction, %d pieces of data have 5G package conversion intention, accounting for about %.4f%%' % (total, 5G, rate))
```

There are 10,000 pieces of data in test set. According to the model prediction, 2,938 pieces of data have 5G package conversion intention, accounting for about 29.38%.

From this, we can see that under the current conditions, the proportion of people who are willing to upgrade the 5g package is low, and more efforts need to be made in the promotion of 5g.

6. Conclusion

1. Box lines, histograms, variance visualization, and other data visualization and analysis tools can help us better analyze customer behavior characteristics, where the split box can make the univariate discrete into N dummy variables after the equivalent of the introduction of non-linearities for the model, which can improve the model expression ability and increase the fit. Reduce the complexity of model operation and improve the speed of model operation.

2. Machine learning models need to be built with good data pre-processing. The handling of missing data is particularly important, so we need to analyze each column of data individually and find the best way to handle it (median, plural, 0, mean or delete).

3. After building different machine learning models, the hyperparameters need to be tuned to find the best parameters for that model and compare the results (accuracy) of the different models. The best model in this project is GradientBoostingClassifier, which we can directly try to use when we encounter similar data.

4. In this project it is possible to clearly determine which users have the idea of upgrading their 5g package. In practice, it can be used to increase targeted advertising or to launch special offers for specific user types. Using these approaches can accelerate the rollout of 5g.

7. Financial management, resource efficiency and resource saving

Research master's thesis is a scientific work related to scientific research, conducting research in order to obtain scientific generalizations, finding principles, and ways to create (modernize) products. Currently, the prospect of scientific research is not determined to scale of the discovery that in the early stages of the life cycle of high-tech and the resource-efficient product is hard enough to reach. Therefore, commercial value is as important as development research. Evaluation of the commercial value (potential) of the development is a necessary condition when searching for sources of funding for scientific research and commercialization of its results. The commercial value is essential for developers who need to understand the state and prospects of ongoing research.

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.

1. Pre-project analysis

Nowadays, the perspective of scientific research is determined not so much by the scale of discovery, which is difficult to estimate at the first stages of the life cycle of a high-tech and resource-efficient product, but by the commercial value of the development. Assessment of the commercial value of the development is a necessary condition when searching for sources of financing for scientific research and commercialization of its results. It is important for developers, who should represent the state and prospects of ongoing scientific research.

It is necessary to understand that the commercial attractiveness of scientific research is determined not only by the excess of technical parameters over previous developments but also by how quickly the developer will be able to find answers to such questions - whether the product will be in demand in the market, what will be its price, what is the budget of the scientific project, how long it will take to enter the market, etc.

The achievement of the goal is ensured by solving the tasks:

- evaluation of the commercial potential and prospects of scientific research;
- identifying possible alternatives to scientific research that meets current resource efficiency and resource conservation requirements;
- research planning;
- resource (resource-saving), financial, budgetary, social, and economic efficiency of research.

2. Competitiveness analysis of technical solutions

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

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

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

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

C - the competitiveness of research or a competitor;

Wi– criterion weight;

Pi – point of i-th criteria.

P_{i1} is China Mobile, P_{i2} is China Unicom

China Mobile and China Unicom are the two largest operator companies in China, their data is more convincing and can also show the advantages of different companies.

Table 1. Evaluation card for comparison of competitive technical solutions

Evaluation criteria	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Ease of operation	0.10	5	5	5	0.50	0.50	0.50
2. Number of business packages	0.13	4	5	3	0.52	0.65	0.39
3. Personal account safety	0,20	4	4	2	0,40	0,40	0,25
4. Smart interface quality	0.10	2	3	4	0.30	0.50	0.50
5. Customer service	0.10	5	7	6	0.36	0.40	0.32
6. Signal coverage	0.07	3	5	4	0.21	0.35	0.28
Economic criteria for performance evaluation							
1. Competitive power	0.10	4	3	5	0.40	0.30	0.50
2. Market penetration rate	0.14	5	6	3	0.20	0.16	0.16
3. Development cost	0.06	3	3	3	0.20	0.10	0.19
Total	**1**	**35**	**41**	**35**	**3.09**	**3.3**	**3.09**

This analysis suggests that the study is effective because it provides acceptable quality results. Further investment in this development can be considered reasonable.

3. SWOT analysis

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

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

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

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

Table 2. SWOT analysis

	Strengths: S1. Build detailed propensity labels for users who are willing to upgrade to 5G S2. Big data analysis can improve business decisions and provide data support.	**Weaknesses:** W1. There is a lot of missing data in the data W2. Too many variables in the data set
Opportunities: O1. User upgrade tendency analysis to facilitate targeted increase in advertising investment. O2. All upgraded user information is visualized,	*Strategy which based on strengths and opportunities: User activation of 5g information is intuitive.*	*Strategy which based on weaknesses and opportunities: Use various data processing methods to deal with missing data and improve the accuracy of the model.*

which is convenient for company operation analysis.		
Threats: This project will calculate user tendencies, and operators should propose targeted strategies.	*Strategy which based on strengths and threats: Provide visual user information.*	*Strategy which based on weaknesses and threats: Different methods used to fill in missing data will lead to different results from data modeling*

4. 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.

This section describes the project stakeholders, the hierarchy of project objectives and the criteria for achieving the objectives.

Project stakeholders refer to individuals or organizations that are actively involved in the project or whose interests may be affected positively or negatively during project implementation or completion. Information about project stakeholders provides in Table 3.

Table 3. Stakeholders of the project

Project stakeholders	**Stakeholder expectations**
The companies where the user opened the business	Easy to use and built model.
Users who have opened services	Objective factors affecting users, age, work, etc.

Table 4. Purpose and results of the project

Purpose of project:	Analyze the behavioral tendencies of 5g users, and predict the number of existing 4g users who will open 5g
Expected results of the project:	Build detailed user information visualization

Criteria for acceptance of the project result:	Modeling and predicting 5G subscribers with an accuracy greater than 90 percent
Requirements for the project result:	1. The project must be completed before May 31, 2022 of the current year.
	2. The results obtained must meet the acceptance criteria of the project results.

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

Table 5. Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours (working days (from table 7) × 6 hours)
1	Supervisor	Head of project	Suggest project direction and review master's thesis.	636 hours
2	Student	Executor	Writing master's dissertations. Through data cleaning, user data analysis, machine learning, result analysis and comparison, modeling correction, and finally predicting the user's 5G activation situation to obtain the accuracy	894hours

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

Table 6. Project limitations

Factors	Limitations / Assumptions
3.1. Project's budget	125000 RUB
3.1.1. Source of financing	TPU
3.2. Project timeline:	1/1/2022 to 30/05/2022
3.2.1. Date of approval of plan of project	1/10/2021
3.2.2. Completion date	16/05/2022

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

Table 7. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
General Technical supervision	26 days(without 5 weekend)	1/10/2021	31/10/2021	Supervisor
Planning project	25 days(without 4 weekend and 1 holiday)	1/11/2021	30/11/2021	Supervisor/ Student
Data Cleaning	29 days(without 5 weekend ,9 holiday and 3 exam day)	1/12/2021	15/01/2021	Supervisor/ Student
User Behavior Analysis	26 days (without 5 weekend)	16/01/2021	15/02/2021	Student
Building models	22 days (without 4 weekend and 2 holiday)	16/02/2021	15/03/2021	Student
Trimming model parameters to increase prediction accuracy	26 days(without 5 weekend)	16/03/2022	15/04/2022	Supervisor/ Student
Preparing of dissertation	21 days(without 5 weekend and 4 holiday)	16/04/2022	16/05/2022	Student

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

Table 8. Gantt chart of Project Schedule

№	Activities	Participants	T_c, days	Duration of the project						
				2021			2022			
				10	11	12	1,2	3	4	5
1	General Technical supervision	Supervisor	26							
2	Planning project	Supervisor /Student	25							
3	Data Cleaninnng	Supervisor /Student	29							
4	User Behavior Analysis	Student	26							
5	Building models	Student	22							
6	Trimming model parameters to increase prediction accuracy	Supervisor / Student	26							
7	Preparing of dissertation	Student	21							

5. 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.

6. Calculation of material costs

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

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

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

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

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

k_T – coefficient taking into account transportation costs.

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

Table 9. Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Electricity of computer	kW/h	894	5.8	5185.2
Internet	Moth	8	350	2800
Papers A4		120	1	120
Pen		2	150	300
Total				8405.2

The personal computer costs 50,000 rubles, the software related to the design model is completely free, and the purchase of the data pays 1,000 rubles to the Chinese data website. So costs of special equipment is 51000 rub.

7. Basic salary

This point includes the basic salary of participants directly involved in the implementation of work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system

The basic salary (S_b) is calculated according to the formula:

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

where S_b – basic salary per participant;

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

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

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v} , \quad (3.4)$$

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

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

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

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

Table 10. The valid annual fund of working time

Working time indicators	
Calendar number of days	365
The number of non-working days - weekend - holidays	 52 14
Loss of working time - vacation - isolation period	48

- sick absence	
The valid annual fund of working time	251

Monthly salary is calculated by formula:

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

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 11. Calculation of the base salaries

Performers	S_{base}, rubles	$k_{premium}$	k_{bonus}	k_{reg}	S_{month}, rub.	W_d, rub.	T_p, work days	W_{base}, rub.
Supervisor	37700	0.3	1.28	1.3	77435.8	2581	106	273586
Student	19200				39360	1312	149	195488
Total:								469074

8. 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:

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

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient (15%);

W_{base} – base salary, rubles.

Additional salary of the supervisor: 41037.9rub

Additional salary of the student: 29323.2rub

9. Labor tax

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

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

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

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 12. Labor tax

	Project leader	**Engineer**
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	273586	195488
Labor tax, rubles	74141.8	52977.2
Total:	127119	

10.Overhead costs

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

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

Overhead is calculated according to the formula:

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

where k_{ov} – overhead rate.

Table 13. Overhead

	Project leader	**Engineer**
Overhead rate	16%	
Salary, rubles	273586	195488
Overhead, rubles	43773.76	31278.8
Total	75051.84	

11.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 14.

Table 14. Items expenses grouping

Name	**Cost, rubles**
1. Material costs	0
2. Equipment costs	51000
3. Basic salary	469074
4. Additional salary	70361.1
5. Labor tax	127119
6. Overhead	75051.84
7. Other direct costs	8405.2
Total planned costs	**801011,14**

12.Evaluation of the comparative effectiveness of the project

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

The integral indicator of the financial efficiency of a scientific study is obtained in the course of estimating the budget for the costs of three (or more) variants of the execution of a scientific study. For this, the largest integral indicator

of the implementation of the technical problem is taken as the calculation base (as the denominator), with which the financial values for all the options are correlated.

The integral financial measure of development is defined as:

(x)

,

where – integral financial measure of development;

C_i – the cost of the i-th version;

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

The obtained value of the integral financial measure of development reflects the corresponding numerical increase in the budget of development costs in times (the value is greater than one), or the corresponding numerical reduction in the cost of development in times (the value is less than one, but greater than zero).

Since the development has one performance, then = 1.

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

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

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

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

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

n – number of comparison parameters.

The calculation of the integral indicator of resource efficiency is presented in the form of table 15.

Table 15 – Evaluation of the performance of the project

Criteria	Weight criterion	Points
1. Ease of operation	0.13	13
2. Noise immunity	0.13	12
3. Safety	0.10	13
4. Smart interface quality	0.10	13
5. Ability to connect to PC	0.08	12
6. Reliability of relay protection	0.07	12
7. Energy efficiency	0.12	15
Economic criteria for performance evaluation		
1. Competitive power	0.10	12
2. Market penetration rate	0.04	13
3. Development cost	0.03	9
4. After-sale service	0.05	13
5. Time to market	0.05	15
Total	1	152

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

$$I_e^p = \frac{I_m^p}{I_f^d} \quad I_e^a = \frac{I_m^a}{I_f^a},$$

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

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

$$E_c = \frac{I_e^p}{I_e^a}$$

.

Thus, the effectiveness of the development is presented in table 16.

Table 16 – Efficiency of development

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

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

13.Conclusion

During the implementation of the financial management section, a comprehensive description and analysis of the financial and economic aspects of the work performed. A list of the work carried out, their performers and the duration of the work stages have been compiled, a line schedule has been drawn up. Also, the cost estimate for the project was calculated, the cost of the project was calculated, the performance indicators of the project were determined and its effectiveness was assessed.

8. Social responsibility

1. Introduction

The developed project aims to use data analysis to visualize user behavior, and use machine learning methods to predict the business opened by users. The development of the program was carried out only with the help of computer.

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

The work was carried out in the hall of residence of TPU (Ucova 13A). Room 206 was a research execution place.

Room characteristics:

- working space width - 2 m, length - 4 m, height - 3 m.
- room area - 8 m2.
- room volume - 34 m3.
- a refrigerator is installed in the room, there is a natural ventilation - exhaust vent, door, window.
- artificial lighting is installed in the room, there is daylight.

2. Legal and organizational issues in providing safety

The regulation of relations between an employee and an employer regarding wages, labor regulations, the specifics of regulating the labor of women, children, people with disabilities, etc., is carried out by the legislation of the Russian Federation, namely the Labor Code of the Russian Federation.

The mode of work and rest provides for the observance of a certain duration of continuous work on a personal computer (PC) and breaks regulated taking into account the duration of the work shift, types and categories of labor activity.

The type of labor activity on a personal computer within the framework of this work corresponds to group B - creative work in the dialogue mode with a PC, category of labor activity - I (up to 2 hours of direct work on a PC).

With an 8-hour work shift and work on a PC that meets the criteria described above, it is necessary to arrange regulated breaks lasting 20 minutes each or lasting 15 minutes every hour of operation.

The duration of continuous work on a PC without a regulated break should not exceed 2 hours.

Effective are unregulated breaks (micropauses) lasting 1-3 minutes. It is advisable to use regulated breaks and micropauses to perform a set of exercises and gymnastics for the eyes, fingers, as well as massage. It is advisable to change sets of exercises after 2-3 weeks.

The duration of the working day should not be less than the time specified in the contract, but not more than 40 hours per week. For employees under 16 years old - no more than 24 hours a week, from 16 to 18 years old and disabled people of groups I and II - no more than 35 hours.

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

Of great importance for the prevention of static physical overload is the proper organization of the workplace of a person working with a PC. The workplace must be organized in accordance with the requirements of standards, specifications and (or) guidelines for labor safety. It must meet the following requirements:

- Provide the possibility of convenient performance of work;
- Take into account the physical severity of the work;
- Take into account the size of the working area and the need to move the worker in it;
- Take into account the technological features of the work process.

Failure to comply with the requirements for the location and layout of the workplace can lead to an employee getting an industrial injury or developing an

occupational disease. The programmer's workplace must comply with the requirements of СанПин 2.2.2/2.4.1340-03.

The design of the equipment and the workplace when performing work in a sitting position should ensure the optimal position of the worker, which is achieved by adjusting the height of the working surface, the height of the seat, the equipment of the space for placing the legs and the height of the footrest.

Figure 1. schematically presents the requirements for the workplace.

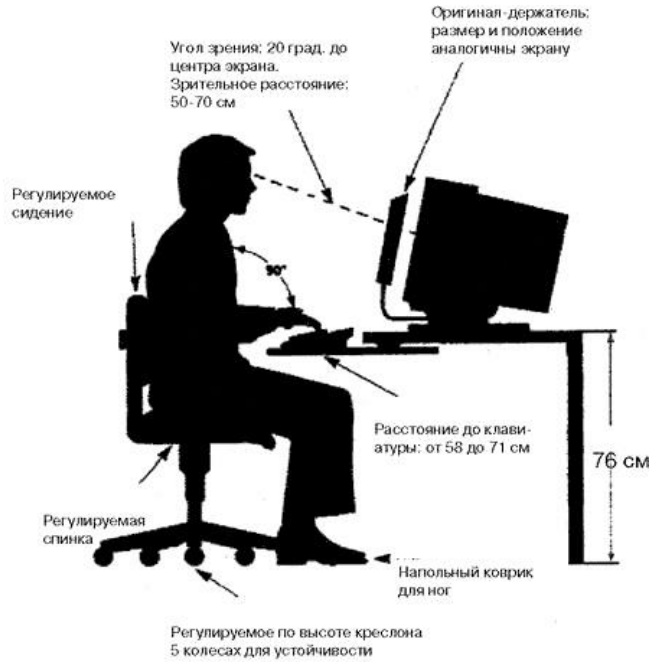


Fig. 1. Workplace organization

Workplace layouts with personal computers should take into account the distances between desktops with monitors: the distance between the side surfaces of the monitors is at least 1.2 m, and the distance between the monitor screen and the back of another monitor is at least 2.0 m. The keyboard should be located on the surface table at a distance of 100-300 mm from the edge facing the user. Fast and accurate reading of information is provided when the screen plane is located below the level of the user's eyes, preferably perpendicular to the normal line of sight (normal line of sight 15 degrees down from the horizontal). Workplaces with a PC when performing creative work that requires significant mental stress or high

concentration of attention are recommended to be isolated from each other by partitions 1.5 - 2.0 m high.

4. Occupational safety

Workplace safety is the responsibility of everyone in the organization.

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

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

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

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

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

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

Table 1 - Potential hazardous and harmful production factors

Factors (**GOST 12.0.003-2015)**	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Excessive levels of noise	+	+	+	**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. Electromagnetic fields	+	+	+	**SanPiN 2.2.4.1329-03** Requirements for protection of personnel from the impact of impulse electromagnetic fields
4. Abnormally high voltage value in the circuit, the closure which may occur through the human body	+	+	+	**Sanitary rules GOST 12.1.038- 82 SSBT.** Electrical safety. Maximum permissible levels of touch voltages and currents.

4.1 Excessive levels of noise

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

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

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

4.2 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 [7].

According to the SanPiN 2.2.2 / 2.4.1340-03 [8] standard, the illumination on the table surface in the area ofthe working document should be 300-500 lux. Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300 lux.

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

4.3 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. Compliance with these standards makes it possible to avoid the negative effects of electromagnetic radiation.

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

Sources of electromagnetic radiation in the workplace are system units and monitors of switched on computers. To bring down exposure to such types of

radiation, it is recommended to use such monitors, the radiation level is reduced, as well as to install protective screens and observe work and rest regimes.

According to the intensity of the electromagnetic field at a distance of 50 cm around the screen along the electrical component should be no more than [9]:

- 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 [10], the design of the computer should provide the power of the exposure dose of x–rays at any point at a distance of 0,05 m from the screen no more than 100 μR/h. Fatigue of the organs of vision can be associated with both insufficient illumination and excessive illumination, as well as with the wrong direction of light.

4.4 Abnormally high voltage value in the circuit

The mechanical action of current on the body is the cause of electrical injuries. Typical types of electric injuries are burns, electric signs, skin metallization, tissue tears, dislocations of joints and bone fractures.

- The following protective equipment can be used as measures to ensure the safety of working with electrical equipment:
- disconnection of voltage from live parts, on which or near to which work will be carried out, and taking measures to ensure the impossibility of applying voltage to the workplace;

- posting of posters indicating the place of work;
- electrical grounding of the housings of all installations through a neutral wire;
- coating of metal surfaces of tools with reliable insulation;
- inaccessibility of current-carrying parts of equipment (the conclusion in the case of electroporation elements, the conclusion in the body of current carrying parts) [11].

5. Ecological safety

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

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

Modern PCs are produced practically without the use of harmful substances hazardous to humans and the environment. Exceptions are batteries for computers and mobile devices. Batteries contain heavy metals, acids and alkalis that can harm the environment by entering the hydrosphere and lithosphere if not properly disposed of. For battery disposal it is necessary to contact special organizations specialized in the reception, disposal and recycling of batteries [12].

Fluorescent lamps used for artificial illumination of workplaces also require special disposal, because they contain from 10 to 70 mg of mercury, which is an extremely dangerous chemical substance and can cause poisoning of living beings, and pollution of the atmosphere, hydrosphere and lithosphere. The service life of such lamps is about 5 years, after which they must be handed over for recycling at

special reception points. Legal entities are required to hand over lamps for recycling and maintain a passport for this type of waste. An additional method to reduce waste is to increase the share of electronic document management [8].

6. Safety in emergency

In the working environment of the PC operator, the following manufactured emergencies may occur [13]:

- Fires and explosions in buildings and communications;
- Collapse of buildings.

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

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

The most typical for the object where the working rooms are located, equipped with a personal computer, the emergency is a fire. Premises for work of PC operators according to the classification system of categories premises for explosion and fire hazard belongs to category D (out of 5 categories A, B, B1-B4, D, D), because applies to premises with non-combustible substances and materials in a cold state[13].

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

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

With a frequency of at least once every three years, it is necessary to measure the insulation resistance of current-carrying parts of power and lighting equipment. The increase in sustainability is achieved through the implementation of appropriate organizational and technical measures, training of personnel to work in emergencies[14].

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

- It is required to stop work, call the fire department by phone "01";
- If possible, take measures to evacuate people and material values;
- Disconnect electrical equipment from the mains;
- Start extinguishing the fire with the available fire extinguishing means;
- Inform the immediate or superior supervisor and notify the surrounding employees;
- In case of a general signal of danger, leave the building in accordance with the "Plan for the evacuation of people in case of fire and other emergencies."

To extinguish a fire, use manual carbon dioxide fire extinguishers (type OU-2, OU-5) located in the office premises, and a fire hydrant internal fire-fighting water supply. They are designed to extinguish the initial fires of various substances and materials, with the exception of substances that burn without air access. Fire extinguishers must be kept in good working order at all times and ready for action. It is strictly forbidden to extinguish fires in office premises using chemical foam fire extinguishers (type OHP-10) [15].

7. Conclusion

Despite the relative simplicity of the design, working space and degree of operation, the considered dangerous and harmful factors can significantly affect the condition and health of the user and the environment. The main points in emergency situations and actions in case of their occurrence were described, as well as legal norms and norms for the operation of the workplace. All these remarks make it possible to use the developed software package effectively and without consequences for employees.

Reference

1. Feature analysis model . Great Cihai reference date 2020-12-01.

2. An Hongwei, Meng Xinna, Gong Lixia, editors; Zhang Aimin, Zhang Lijuan, Zheng Lifang, Zhang Guoqiang, deputy editors . Probability Theory and Mathematical Statistics (for independent colleges) : China Railway Press , 2016.01.

3.Data Analysis/MiningHow to deal with categorical features? Common encoding method? Python implementation -

Link: https://blog.csdn.net/qq_41595507/article/details/112095732.

4. Gradient Boosting Algorithm

Link - https://zhuanlan.zhihu.com/p/86354141

5. Minister of Industry and Information Technology Miao Wei attended the first meeting of the IMT-2020 (5G) promotion group and issued a letter of appointment to the group leader, President Cao Shumin--China Academy of Information and Communications Technology [citation date 2021-05-15]

6. The first global 5G conference was held. Chinese government network reference date 2021-05-15

7. GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements.

8. SP 2.4.3648-20. Sanitary and Epidemiological Requirements for Organizations of Education and Training, Recreation and Recreation of Children and Youth

9. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.

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

11. SanPiN 2.2.4.1329-03 Requirements for protection of personnel from the impact of impulse electromagnetic fields

12. GOST 12.1.038-82 Occupational safety standards system. Electrical safety

13. Federal Law "On the Fundamentals of Labor Protection in the Russian Federation" of 17.07.99 № 181 – FZ

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

15. GOST R 52105-2003 Resources saving. Waste treatment.