**Personalized recommendation of Intelligent Service for users in the telecommunications industry**

**Members：** 张乐乐 201828013229014 计算技术研究所

童旭升 201828013229043 计算技术研究所

程晓翠 201828012637004 南京土壤研究所

李俊杰 201828012638011 南京土壤研究所

# Introduction

As one of the national basic industries, the telecommunications industry covers a wide range of users and is particularly important in supporting national construction and development. With the rapid development and popularization of Internet technology, the traffic consumed by users has also become a blowout trend. In recent years, telecom operators have launched a large number of telecom data plan to meet the differentiated needs of users. In the face of a wide variety of cellphone data plan, how to choose the most suitable One is very important for operators and users, especially in the context of the slowdown in the telecommunications market and the increasingly fierce competition for stock users. According to the personalized recommendation problem of telecommunication phone plan, a personalized recommendation model of telecommunication phone plan based on user consumption behavior is constructed by data mining technology. According to the result of user business behavior, the user's consumption habits and preferences are analyzed, and the most suitable cellphone plan is matched. The purpose is to improve user perception and drive user demand, so as to achieve the goal of user value improvement.

Personalized recommendation of the phone plan, using the existing user attributes (such as personal basic information, user portrait information, etc.), terminal attributes (such as terminal brands, etc.), business attributes, consumption habits and preferences to match the most appropriate phone plan for users, and complete subsequent personalized services. In short, there are two main problems need to solved: the information overload problem and the user's purposeless search problem. The various phone plans meet needs of users when they have a clear purpose. However, personalized recommendations can help them discover new content of interest when they have no clear purpose.

# Summary

Intelligent forecasting of telecom services is a very important issue for telecom operators and users. According to the results of user business behaviors, analyzing user consumption habits and preferences, matching the most appropriate services for users, which can enhance user perception and drive user demand, and also can bring operator to a steady stream of profits. This project mainly uses China Unicom's real data set for data mining, machine learning and other technologies to study the relationship between various characteristics of users and current service, and predict the current service of test set users. We first cleaned and visually analyzed all the data using pandas, numpy, matplotlib, etc. Then we constructed two different feature models for processing and using LightGBM, decision tree, random forest, logistic regression and other algorithms in the scikit-learn framework. The two features were predicted, and finally the best results (feature\_2 with LightGBM) were continuously optimized. We proposed the 10+2 model and 8+3 models to get further optimization, The LightGBM algorithm was continuously adjusted to improve the prediction performance. The final 8+3model The F1 score is 91.55%. This model provides a good reference for operators to personalize the recommended telecommunication services.

# Techniques

**Python:**

Numpy, pandas, matplotlib, scikit-learn

**Data Mining(Machine Learning):**

LightGBM, Decision Tree, Random Forest, Logistic Regression

# Solution Explanation

According to the above characteristics, we use the decision tree, random forest and Lightgbm algorithms to predict the telecommunication packages that users may purchase.

## 1. decision tree

Decision tree is a non-parametric supervised learning method. Its essence is a conditional judgment statement, which is mainly used for classification and regression. The algorithm predicts the target value based on data features and learning decision rules through a judgment mechanism such like if-else.

## 2. Random forest

Random forest refers to a method of training, classifying and predicting sample data using multiple decision trees. There is no correlation between each decision tree in the random forest, so when there is a new sample to be classified, the final classification result is determined by the vote of the multi-tree classifier. The grid search and cross-validation methods are used to determine the parameter estimators, which represents the number of trees established.

## 3. LightGBM

LightGBM is a gradient Boosting framework that uses a decision tree-based learning algorithm. It can be said to be distributed and efficient, with the following advantages: faster training efficiency, low memory usage, higher accuracy, support parallel learning, can handle large-scale data.

# Management Implementation

## Work arrangement

Zhang Le: The responsible project content mainly includes method design, code writing, and report writing.

Tong Xu Sheng: The responsible project content mainly includes method design, code writing, and report writing.

Cheng Xiaocui: Responsible for the final ppt production of the project.

Li Junjie: Responsible for the writing of the report, the modification of ppt.

## Project schedule

First stage（2019/3/22-2019/4/10）：After the teachers are grouped, the first stage of the task is to become familiar with the members of the team, to understand the professional background of the members. Additionally, we need to determine a theme for this project. We have chosen the smart phone plans personalized matching for the telecom industry stock users as our team's project.

Second stage（2019/4/11-2019/4/19）：At this phrase we will convert the datasets into the testing sets and the training sets. The data training set of this project mainly has 27 dimensions, about 543,990 data records, and the test set of the project contains 27 dimensions, and about 200,000 data records. Cleaning and pre-processing data is an essential step in the entire project before the project begins, which can help us better understand the distribution of the data, the connections between the variables and the meaning behind the data.

Third stage (2019/4/20 – 2019/5/20): During this period, our team's main job is to use the LightGBM model to analyze the prediction results of the full features, analyze the accuracy of the 11 plans results, and finally imagine designing a new model to transform the 11 classification problems into 10 classification problems. Then, the two classification models of 99998826 and 99999827 are separately predicted, which is our own modified 10+2 model. In addition, we transform the 11 classification problems into 8 classification problems based on 4G services and 3 classification problem based on Internet services, and get the 8+3 model. We use the further processed Feature\_1 and Feature\_2 to perform operations under the decision tree, random forest, and lightGBM algorithm respectively, and find that the best performance can be achieved under lightGBM.

Fourth（2019/5/21 – 2019/6/3）：

At this stage，Zhang Le and Tong Xusheng are more familiar with the code, so the model establishment and feature engineering in the report are completed by them. Li Junjie and Cheng Xiaocui are responsible for the writing of the report and the production of ppt.

# Data Analysis

## 1. Data Cleaning

The data training set of this project mainly has 27 dimensions, about 543,990 data records, and the test set of the project contains 27 dimensions, and about 200,000 data records. The content mainly includes user attributes (personal basics information, etc.), terminal attributes (such as terminal brands, etc.), business attributes, consumption habits. The relevant data headers are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 中文名 | 数据类型 | 说明 |
| USERID | 用户ID | VARCHAR2(50) | 用户编码，标识用户的唯一字段 |
| current\_type | 套餐 | VARCHAR2(500) | / |
| service\_type | 套餐类型 | VARCHAR2(10) | 0：23G融合，1：2I2C，2：2G，3：3G，4：4G |
| is\_mix\_service | 是否固移融合套餐 | VARCHAR2(10) | 1.是 0.否 |
| online\_time | 在网时长 | VARCHAR2(50) | / |
| 1\_total\_fee | 当月总出账金额\_月 | NUMBER | 单位：元 |
| 2\_total\_fee | 当月前1月总出账金额\_月 | NUMBER | 单位：元 |
| 3\_total\_fee | 当月前2月总出账金额\_月 | NUMBER | 单位：元 |
| 4\_total\_fee | 当月前3月总出账金额\_月 | NUMBER | 单位：元 |
| month\_traffic | 当月累计-流量 | NUMBER | 单位：MB |
| many\_over\_bill | 连续超套 | VARCHAR2(500) | 1-是，0-否 |
| contract\_type | 合约类型 | VARCHAR2(500) | ZBG\_DIM.DIM\_CBSS\_ACTIVITY\_TYPE |
| contract\_time | 合约时长 | VARCHAR2(500) | / |
| is\_promise\_low\_consume | 是否承诺低消用户 | VARCHAR2(500) | 1.是 0.否 |
| net\_service | 网络口径用户 | VARCHAR2(500) | 20AAAAAA-2G |
| pay\_times | 交费次数 | NUMBER | 单位：次 |
| pay\_num | 交费金额 | NUMBER | 单位：元 |
| last\_month\_traffic | 上月结转流量 | NUMBER | 单位：MB |
| local\_trafffic\_month | 月累计-本地数据流量 | NUMBER | 单位：MB |
| local\_caller\_time | 本地语音主叫通话时长 | NUMBER | 单位：分钟 |
| service1\_caller\_time | 套外主叫通话时长 | NUMBER | 单位：分钟 |
| service2\_caller\_time | Service2\_caller\_time | NUMBER | 单位：分钟 |
| gender | 性别 | varchar2(100) | 01.男 02女 |
| age | 年龄 | varchar2(100) | / |
| complaint\_level | 投诉重要性 | VARCHAR2(1000) | 1：普通，2：重要，3：重大 |
| former\_complaint\_num | 交费金历史投诉总量 | NUMBER | 单位：次 |

Cleaning and pre-processing data is an essential step in the entire project before the project begins, which can prevent the subsequent data mining from going wrong. The data cleaning process is mainly divided into the following steps:

1. Filter the non-null values of the two categories of gender and age in the training set and test set, and then convert these two variables into integer.
2. Find the non-null values in the numerical variables such as '2\_total\_fee' and '3\_total\_fee' in the training set, and then convert these two variables into float.
3. Find out the null values in the numerical variables such as '2\_total\_fee' and '3\_total\_fee' in the test set, and replace 0 to the null value, then convert the two variables into float.
4. we merge the data, re-index, and zero the null value in the entire data set.

## 2. Exploratory Data Analysis

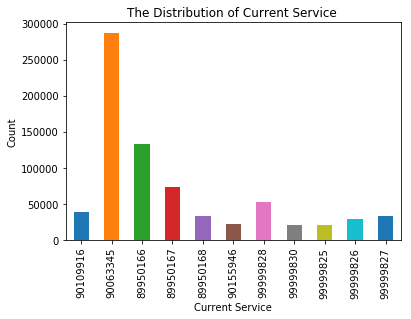
### 2.1 Discrete Variable Analysis

After the data cleaning, the Exploratory Data Analysis can help us better understand the distribution of the data, the connections between the variables and the meaning behind the data. We conduct exploratory data analysis on discrete and continuous variables based on whether the data is continuous or not. Firstly, we describe the distribution of user’s gender, and we want to explore more information about the data (the zero value in the figure represents the outliers, here take the primal value):

****

**Figure 1 the distribution of the user gender**

The frequency of each plan was displayed using a columnar figure, which allowed us to better understand the distribution of cellphone plans:

****

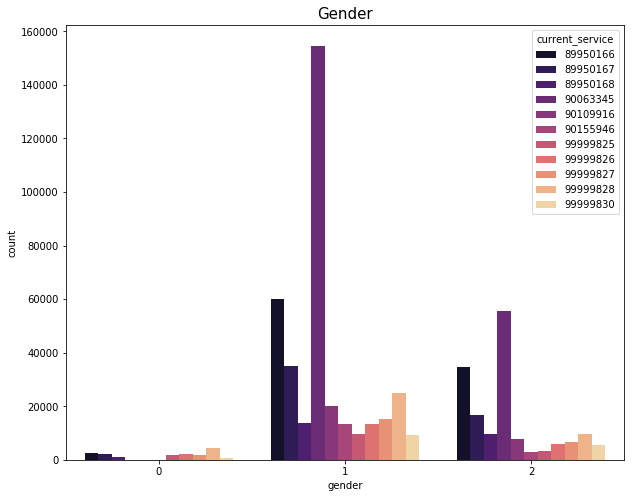
**Figure 2 Frequency statistics for 11 types of phone plans**

After we understand the distribution of these 11 phone plans, we need to explore what the 11 cellphone plans mean.

|  |  |
| --- | --- |
| Service | Name |
| 99999826 | 4G 196RMB |
| 99999828 | 4G 136RMB |
| 89950166 | 4G 76RMB |
| 99999827 | 4G 166RMB |
| 89950167 | 4G 106RMB |
| 99999830 | 4G 76RMB |
| 90109916 | Ant Card |
| 89950168 | 4G 56RMB |
| 99999825 | 4G 296RMB |
| 90155946 | Tencent Card(level 2) |
| 90063345 | Tencent Card(level 1) |

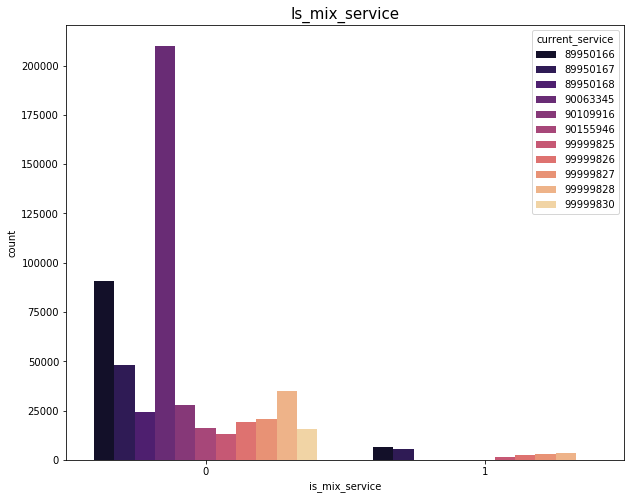
**Figure 3 The meaning of the 11 kinds of services**

After doing the above visual analysis, we found an interesting phenomenon by linking the type of the plans to the gender. The population of the men in the same cellphone plan is more than the women, and the 90063345 plan is more popular than other cellphone data plans.



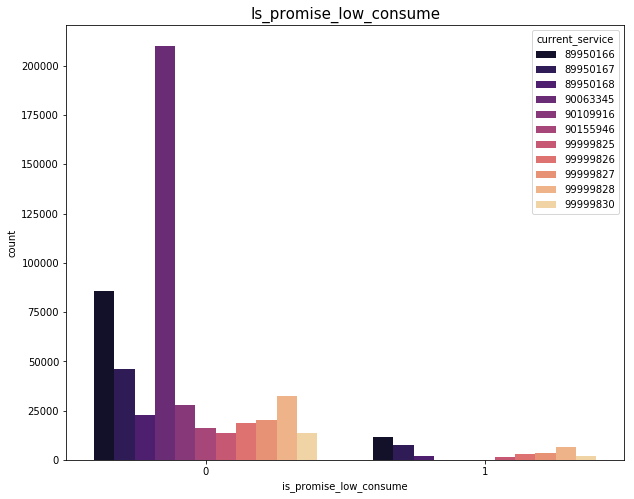
**Figure 4 Gender distribution of different phone plan types**

In addition to the above analysis, it is necessary to display more specific information about the plans. For example, we need to know the proportion of the mix-service phone plans in the total plans (it can be seen that most users are not using the mix-service cellphone plans):



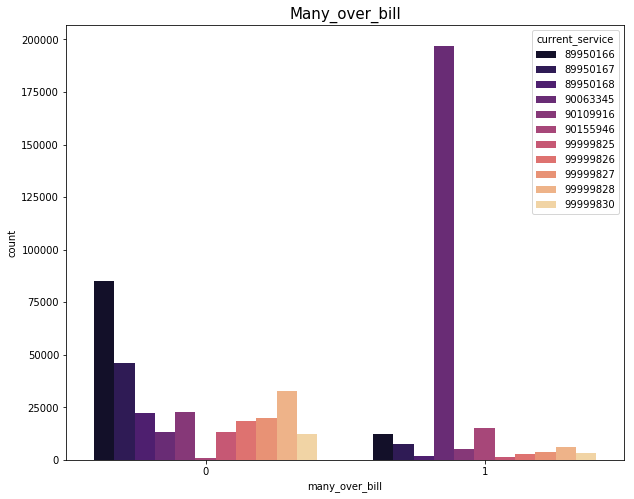
**Figure 5 the distribution of the mix-service**

Furthermore, we need to examine whether the user promises the low consuming. Analysis of the user's low consumption can help us to more effectively determine the habits of users to use which cellphone plan in the future. Therefore, it is very important to visually analyze the user's low consumption situation.



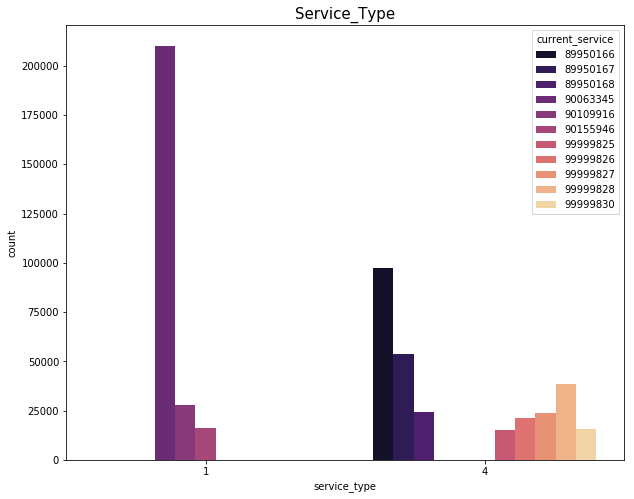
**Figure 6 The distribution of the low consuming**

Of course, according to the over bills, it can also help us analyze the habits of user about the phone plan usage, and predict the use of the plan. In the figure, 0 means not over bill, 1 means over bill, basically each cellphone plan has over bills, but the 90063345 plan has 90% of users are the over bill, which is mainly due to the characteristics of Tencent King card. Users generally exceed the usage of the plan, and this situation can be well explained in a realistic sense.

****

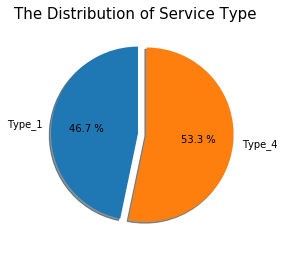
**Figure 7 The characteristic of over-bill user**

After visualizing the 11 specific cellphone plan types, we need to analyze the meanings of type1 and type2 respectively. In the analysis chart, it can be found that 0 represents the fusion of 2, 3G; 1 refers to 2I2C; 2 refers to 2G; 3 means 3G; 4 represents 4G. Especially, 2I2C refers to the cooperation project with Internet companies and tens of millions of users, which is fully operated by e-commerce mode. Therefore, type1 and type2 represent the 2I2C and 4G network types, respectively.



**Figure 8 The distribution of service-type**

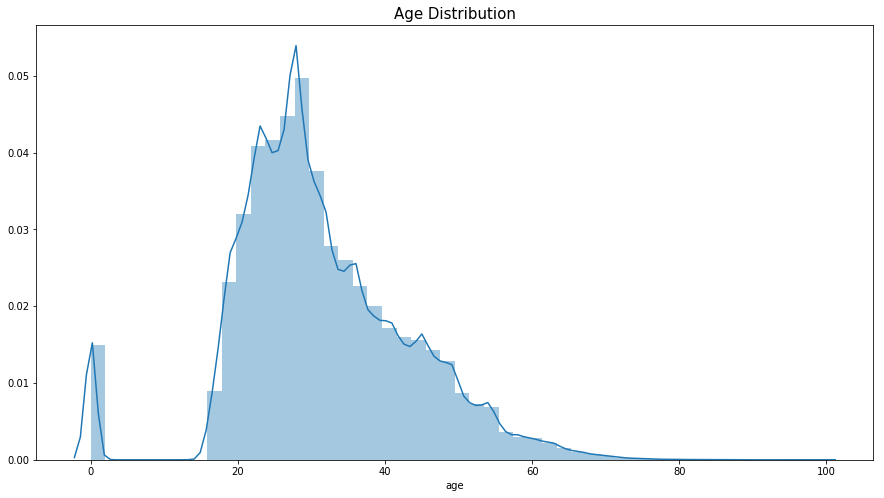
We can continue to count the frequency of the plans after determining the two types of cellphone plans, so we have a pie chart for the frequency statistics of the Internet phone plan and the normal Unicom plan:



**Figure 9 Frequency statistics for two types of plan**

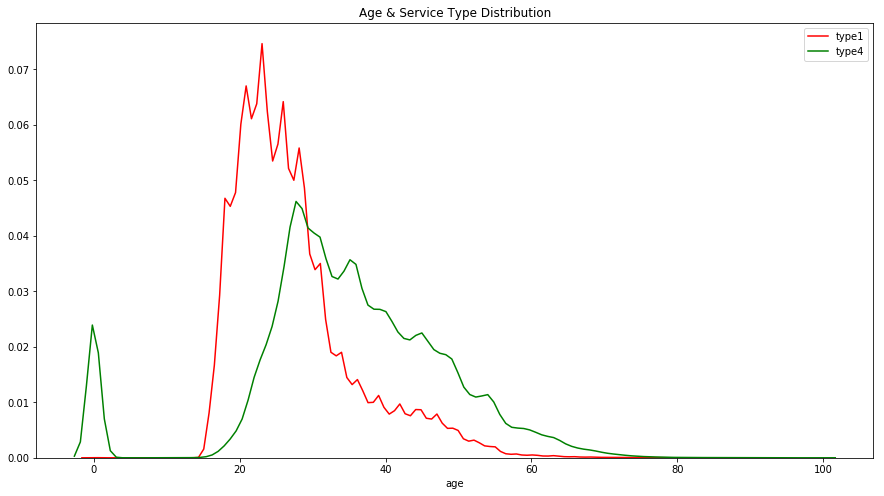
### 2.2 Continuous Variable Analysis

Firstly, we describe the distribution of user age. Figure 10 shows that most users are mainly distributed in the youth group and middle-aged group, which basically conforms to the distribution of telecom user groups, but there are many abnormal values of 0 years old in the figure. For the outliers, we tried to replace the mean corresponding to the service-type field and use the original value, and finally we selected the original value.



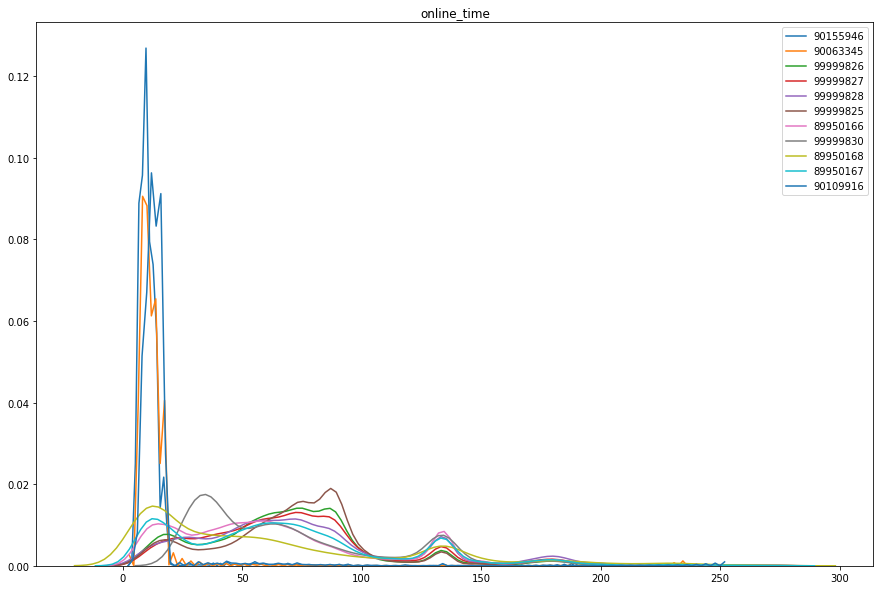
**Figure 10 the distribution of the user age**

In order to further confirm the relation between the service types and ages, we select their distribution in different age groups for visual analysis. As can be seen from Figure 11, the 2I2C network types are mostly concentrated in teenagers, and the 4G network types are concentrated on middle age.



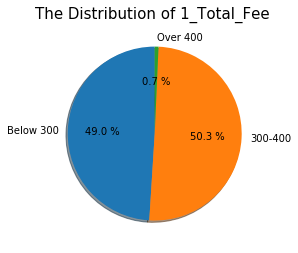
**Figure 11 Distribution of two types of plans at different ages**

Nowadays, the impact of the Internet on human society has penetrated into every aspect. In order to better understand the habits of users, it is essential to understand the length of time they spend on Internet. Therefore, our team chooses the time of online as a judgment for different plan types.



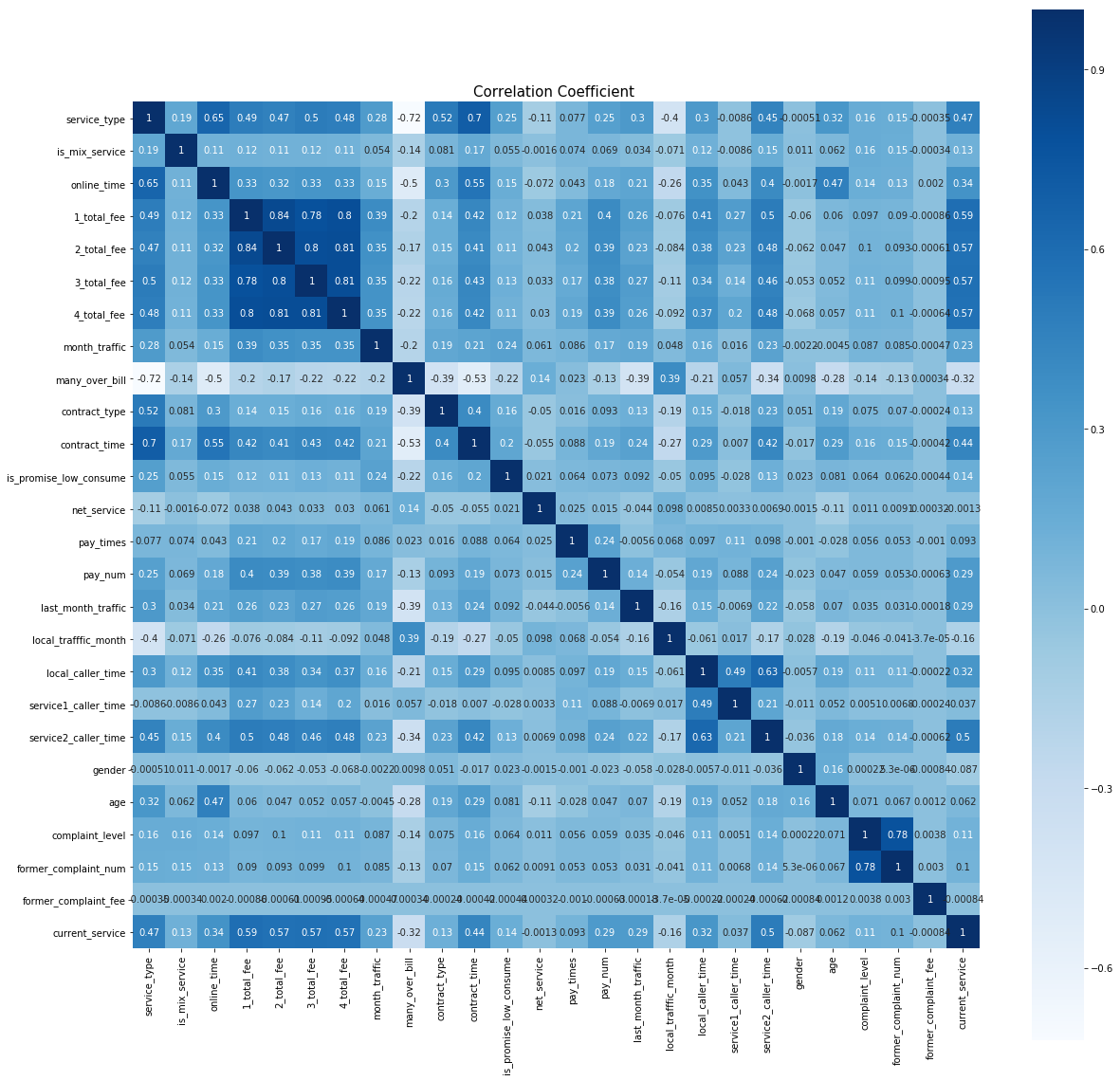
**Figure 12 the change of the time spent on the internet**

The habit of user using the phone plan is inevitably very closely related to the user's payment. Therefore, understanding the payment situation of the user can well match the suitable plan for themselves in the future. We have displayed a pie chart for the payment amount at different stages, which is a good understanding for the user's payment.



**Figure 13 The situation of the user’s payment**

Finally, in the data preprocessing process, the most important thing is to explore the relationship between variables and which variables are most relevant to current-service, so we have a correlation analysis of all variables, Figure 13 can clearly see the relationship between variables:



**Figure 14 The correlation between the different variables**

It can be seen from the above figure that there is a strong correlation between total-fee, and current-service. After the correlation visualization analysis, we also found the correlation between different variables and current-service, and got the same answer. Current-service has the strongest correlation with total-fee.

The steps of correlation analysis can help us to understand which factors have the greatest impact on the current-services. In the following model construction, the above steps have great reference significance, and the data preprocessing and visual analysis play an irreplaceable role in the project.

## 3. Feature engineering

### 3.1 Feature\_1

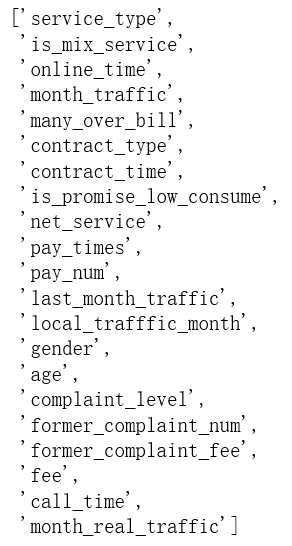
The continuity variables, age and online\_time are discretized;

The first four total\_fee with high correlation with the target value are averaged to form a new feature *fee*;

the local\_caller\_time, service1\_caller\_time, service2\_caller\_time are summed to form a new feature *call\_time*;

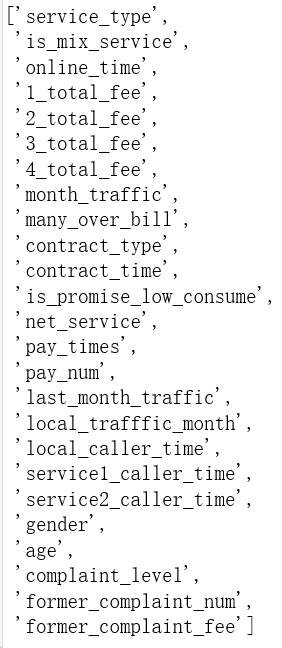
Use last\_month\_traffic to subtract the last\_month\_traffic to get the traffic in the package actually used this month *month\_real\_traffic*.

Then we get feature\_1 as follows:



### 3.2 Feature\_2

Remove uesr\_id from the original feature:



## 4. Classification model and preliminary results

According to the above characteristics, we use the decision tree, random forest and Lightgbm algorithms to predict the telecommunication packages that users may purchase.

### 4.1 decision tree

The decision tree is a nonparametric supervised learning method used primarily for classification and regression. The goal of the algorithm is to create a model that predicts the target value by inferring the data characteristics and learning the decision rules.

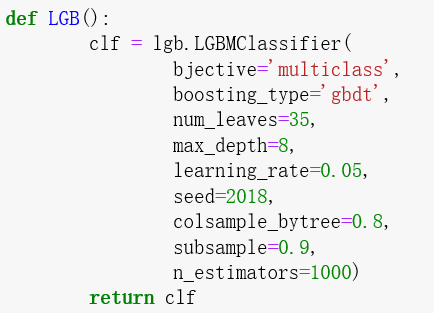
### 4.2 Random forest

Random forest refers to a method of training, classifying and predicting sample data using multiple decision trees. For the classification problem, the final classification result is determined by voting by the multi-tree classifier. The grid search and cross-validation methods are used to determine the parameter n\_estimators, which represents the number of trees established.

### 4.3 LightGBM

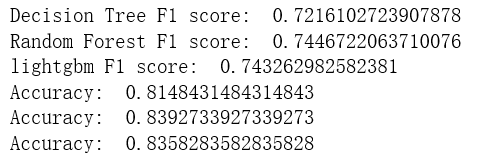
LightGBM is a gradient Boosting framework that uses a decision tree-based learning algorithm. It can be said to be distributed and efficient, with the following advantages: faster training efficiency, low memory usage, higher accuracy, support parallel learning, can handle large-scale data.

Use grid search to determine the parameters as follows:

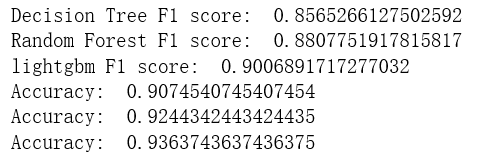


**preliminary results:**

The feature\_1 is input into each model for training, and the following results are obtained:



The feature\_2 is input into each model for training, and the following results are obtained:



## 5. Further optimization and analysis

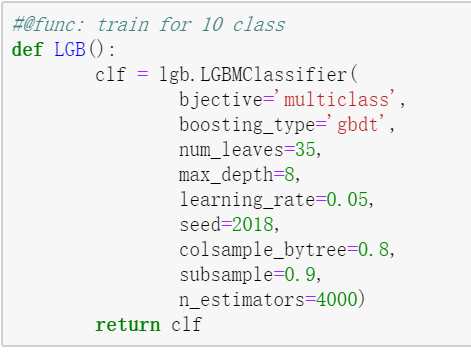
### 5.1 10+2 Model

Firstly, we analyzed the prediction results of Feature\_1 with LightGBM algorithm, and calculated the accuracy of 11 kinds of service prediction. The results are as follows:

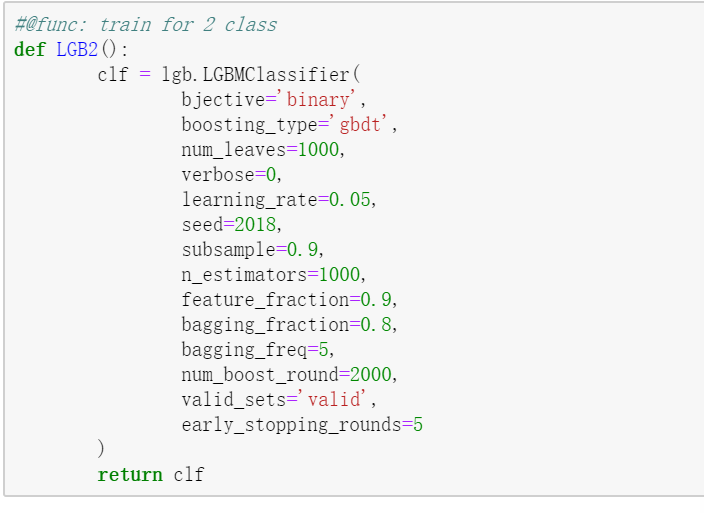
|  |  |  |
| --- | --- | --- |
|  | **Service** | **Accuracy** |
| **0** | 89950166 | 0.8650799884795644 |
| **1** | 89950167 | 0.8178586816027574 |
| **2** | 89950168 | 0.9491994703262309 |
| **3** | 90063345 | 0.9953289823669085 |
| **4** | 90109916 | 0.9919016489413601 |
| **5** | 90155946 | 0.9446532285616672 |
| **6** | 99999825 | 0.8713022781366881 |
| **7** | **99999826** | **0.6847688227833506** |
| **8** | **99999827** | **0.6501110202173659** |
| **9** | 99999828 | 0.7401551312649165 |
| **10** | 99999830 | 0.7713903743315508 |

It can be found that the two services of 99998826 and 99999827 have the lowest prediction accuracy rates of 68.48% and 65.01% respectively. It is found that often distinguishes errors in real, so we envisage to design a new model, which will consider both 99998926 and 99999827 to be 99998826 first (because The accuracy of the model is higher for 99999926.) At this time, the 11-classification problem is transformed into the 10-classification problem. Then we designed the two classification models of 99998826 and 99998827 separately for prediction. Finally, the results of the two models can be integrated to obtain the final forecast result.

Based on the previous predictions, lightGBM is very effective in solving this multi-classification problem, so the integrated 10-classification problem still to use lightGBM for prediction under feature\_1, using grid search optimization parameters, the key parameters are as follows:



For the two-classification problem, we also adjusted the parameters of lightGBM as shown in the figure below, and use the logistic regression commonly used in the two-class to predict (using the default parameters), and finally compare the results.



The overall processing flow is:

1) LightGBM:

**Divide the result by 99999826 to Res\_1 and Res\_2**

**Data Processing Optimize LightGBM**

**Predict test data**

**Replace 99999827 to 99999826**

**(10 Class)**

**Integrate Res\_1 Res\_3**

**and Evaluation**

**Training Data**

**Select the recode of 99999827 and 99999826**

**(2 Class)**

**Utilize Res\_1 as test data to predict (Res\_3)**

**Data Processing Optimize LightGBM**

2) LightGBM+Logistic Regression:

**Divide the result by 99999826 to Res\_1 and Res\_2**

**Data Processing Optimize LightGBM**

**Predict test data**

**Replace 99999827 to 99999826**

**(10 Class)**

**Integrate Res\_1 Res\_3**

**and Evaluation**

**Training Data**

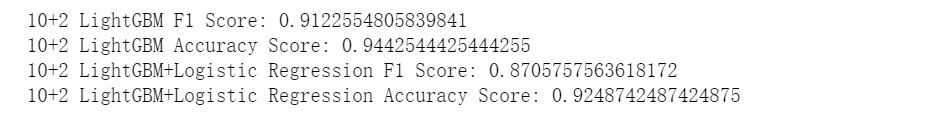
**Utilize Res\_1 as test data to predict (Res\_3)**

**Select the recode of 99999827 and 99999826**

**(2 Class)**

**Data Processing Optimize Logistic Regression**

After the results of the two models were integrated, the classification results were finally evaluated using the accuracy and multi-class F1 Score.



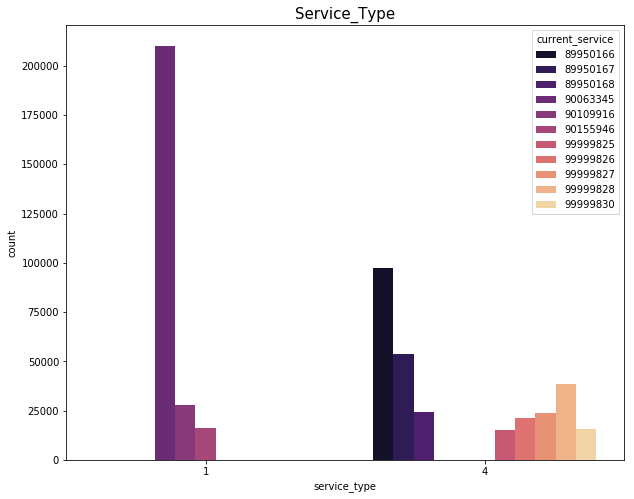
**compare results**

The classification performance of the LightGBM algorithm under the 10+2 model is much higher than that of LightGBM+Logistic Regression.

The LightGBM algorithm is 1.22% more efficient than the direct 11 classification in the 10+2 model.

### 5.2 8+3 model

After deep analysis of the visual data, it was found that the service\_type feature is very useful, and the user took only 1 or 4 under this attribute.



Further review found that 1 represents 2I2C business type 4 represents 4G business type, and 2I2C business type is the type of service jointly launched by China Unicom and the Internet company, such as Tencent King Card.

Based on the above foundation, we transform the 11 classification problem into an 8 classification problem based on 4G services and a 3 classification problem based on Internet services.

The overall processing flow is:

**4G**

**(8 Class)**

**Data process Optimize LightGBM**

**Predict**

**Train&Test Data**

**Integrate the result and Evaluation**

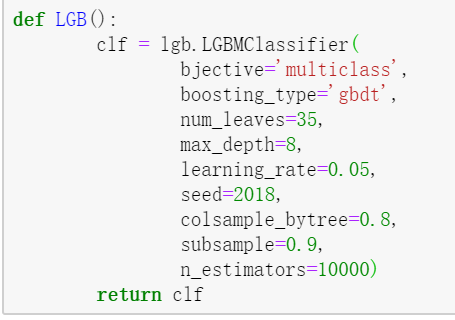
**Internet**

**(3 Class)**

**Data process Optimize LightGBM**

**Predict**

According to the previous experience, LightGBM is used here to perform grid search optimization parameters. The key parameters are as follows:



Similarly, after the results of the two models were integrated, the classification results were finally evaluated using the accuracy and multi-class F1 Score.

C:\Users\zhangle\AppData\Local\Temp\1559204638(1).png

**compare results**

The 8+3 model LightGBM algorithm achieved better performance than the 10+2 model, and F1 increased from 91.226% to 91.593%.

# Conclusion

From the Feature\_1 obtained by processing the feature and the Feature\_2 removing the User\_id, the results obtained by the decision tree, the random forest, and the lightGBM algorithm are respectively analyzed. Feature\_2 can achieve the best performance under lightGBM(F1 Score is 90.069%, Accuracy is 93.637%).

On this basis, in order to further improve the performance, we analyzed the best prediction results (Feature\_2, LightGBM) of the initial processing and proposed a 10+2 model. After verification, the performance has been improved. Further, we analyzed the features to propose the 8+3 model. The experiment showed that the performance was better than the 10+2 model, the F1 Score was 91.593%, and the accuracy was 94.667%.