1. Background

In recent years, 5G networks has becoming a powerful support for the construction of smart cities. It is important to identify potential 5G users in 4G users. Based on the monthly user data of 5G package replacement, we analyze the behavioral characteristics of 4G users replacing 5G packages and build a 5G package latent customer identification model from the basic information, consumption behavior, broadband and other information to identify the current 4G users with the demand group of replacing 5G packages and carry out 5G latent customer marketing.

The purpose of this project is to build a potential customer identification model for 5G packages based on existing data to predict whether 4G users will switch to 5G services. The tools we used are Python3 and Jupyter Notebook.

2. Data Overview

Datasets consists of 60000 users' information in sales area A and B, including gender, age, VIP class, market segments, ARPU, DOU, MOU, balance arrears, data utilization, terminal types, whether expect to have 5G packages and other information.

3. Exploratory Data Analysis

3.1 Data Merge

We have data sets ***train_set.csv*** and ***train_label.csv*** for model training, their information is shown below:

name	data type	column	row	information
train_label.csv	int64（1） object（1）	2	140000	5G enabled label 1：enabled 0：not enabled
train_set.csv	float64(32) int64(10) object(3)	45		User Features

Table1: dataset information

For the indexed data, it is observed that ***user_id*** corresponds to ***product no*** and the label data uses user_id, so user_id is selected as the data index and the product_no is discarded. So, we merge ***train_label.csv*** into ***train_set.csv.***

3.2 Missing value analysis

Tables have 14 features with no missing values, most of them are in the user information group as well as

the user behaviors group, in particular, there is no data missing in the user identification dimension.

There are many features that have the same number of missing values, and we learn from that all the features in the **consumer behavior information dimension** are missing 89 pieces of data, considering whether 89 users have missing. All features in the **Arrears dimension** are missing 392 data, consider whether 392 users have missing value in Arrears and whether they are the same user with missing value in the consumer behavior information dimension. All features in the **Signing Tag dimension** are missing 774 data and features in the **package dimension** are missing 6,617 pieces of data, both of them are considering the same as the Arrears dimension.

Finally, with the sorted data we found a number of features with high missing number. In the user base information dimension, there are 6849 missing values in VIP Class. In the broadband information dimension, there are 101,060 missing features for Broadband Bandwidth and Broadband Bandwidth Activation Status, which is a serious missing feature. Among the other information features, the 5G Data feature is indeed missing 132,559 data, which is very serious, whether to delete this feature needs further analysis.

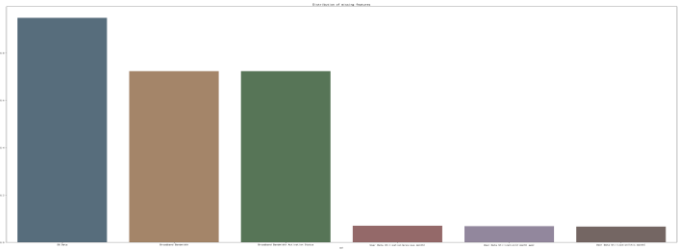

Figure1: Distribution of missing features

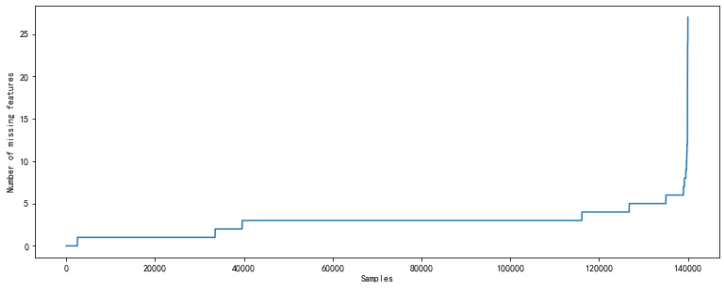

Figure2: Sample Distribution of Missing Features

From Figure 1, we can observe that there is one feature with a missing proportion of more than 0.8, and three features with a missing proportion more than 0.6 but less than 0.8. From Figure 2, the data shows a phased absence, and we can initially determine that multiple missing features in the same dimension are actually the same sample. Some of the samples have more missing features, so we can consider deleting this part of the sample Subsequently.

3.3 Data category analysis

There are two common types of data: categorical features and numerical features. ***Pandas*** will automatically declare the data category when it reads the data, resulting in some of the desensitized category data being perceived as numerical data.

For example, the terminal type, because the data needs to be desensitized, the data values of 1, 2.... In essence, they are all category features, the original data for terminal type which could be Huawei, Samsung, Apple, etc. Here we use the same-value filtering method to classify continuous and categorical variables in numerical variables, which means that if the number of different values in a numeric feature identified by pandas is larger than 10, then the feature is a numeric feature, otherwise it is a category feature.

3.4 Category-feature independent analysis

The single feature independent analysis is mainly for category-based features and numerical features to be analyzed one by one, in the previous analysis it can be found that the number of users does the information within each dimensional group is relatively consistent, so the features in the same dimension can also be compared and analyzed at the same time:

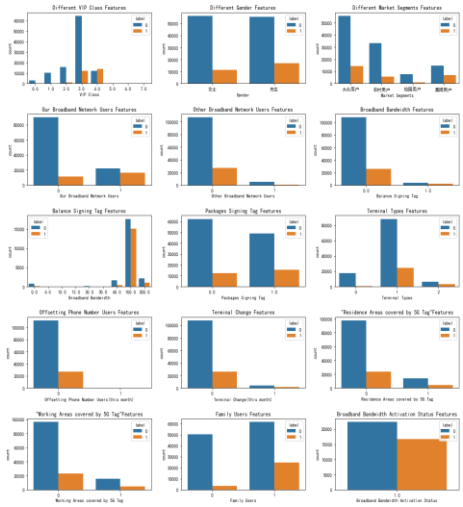

Figure3: Overall single category feature analysis

According to the figure above, the following valid information can be extracted:

1. Among genders, male 5G users have a slightly higher turn-up rate than females. **2.**In the market segment, company users are more likely to turn on 5G. **3.** Among VIP Class, four-star users are more likely to use 5G. **4.** Broadband Network Users are more inclined to use 5G than other Broadband Network Users. **5.** The percentage of users with a broadband bandwidth of 100 is higher and 200 is the second highest. **6.** Family users are more likely to use 5G than non-home users. **7.** The highest number of users with terminal type1 and the highest rate of 5G users with terminal type2. **8.** the broadband activation feature has only unique values and missing values.

3.5 Numerical-feature independent analysis

For the remaining numerical features, there are two exceptions to the numerical features that need to be analyzed separately: **age** and **Time in Network**.

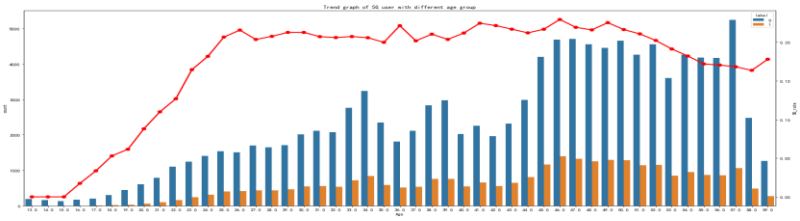

Figure4: Trend graph of 5G users with different age groups

Figure5: Trend graph of 5G users with different Time in-network lengths

From Figure 4, the overall percentage of 5G users between the ages of 25-50 is higher, basically ranging above 20%, the proportion of 5G users between 13-25 years old is on an upward trend with age, the proportion of 5G users above 50 years old tends to decrease with age.

In Figure 5, the sample with the highest percentage of time on the network is 1, but also the lowest percentage of 5G users. When the length of Time in-network is between 2 and 10, the proportion of 5G users ranges from 25% to 30%. When the length of Time in-network is larger than 10, the proportion of 5G users is higher than 30%, and shows an accelerating upward trend.

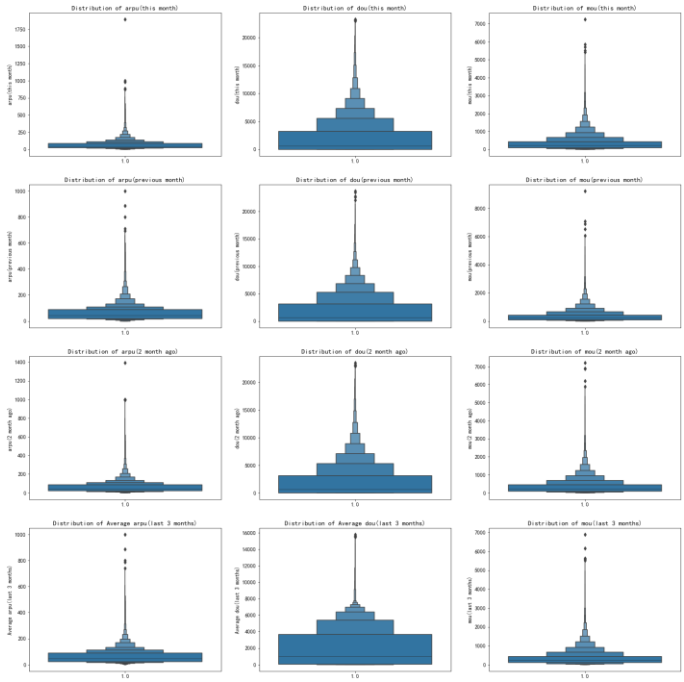

Figure6: Numerical features distribution

Finally, we plot the distribution, analyze the rest of the numerical features, consider subsequent filtering of outliers for extreme minima, and dimensionless the data so that the different features remain on a single scale

and the model is trained more efficiently.

4. Data cleaning and missing value replacement

4.1 Feature Combination

arpu dou mou can use only the average of three months to represent, we can delete the data of the column. For the same reason, similar dimensions can be replaced by the average column.

Again, we define a function to find missing values, and by counting the columns we find a large number of missing values. We simply fill each column first by traversing it, filling in the missing values for numeric variables by the median, and filling in the missing values for categorical variables by the plural.

4.2 Missing value handling

We use Mode to fill Categorical Variable, use Median to fill Numerical Variable. For ***Broadband Bandwidth*** and ***5G Data,*** there are only two situations, so we use 0 to fill in the missing values.

4.3 Categorical variable encoding

Here we use the more widely applicable ***Label encoder*** to encode the values of the various features, with the following results:

	User ID	User Number	Gender	Age	VIP Class	Time in Network	Market Segments	Average arpu(last 3 months)	Average dou(last 3 months)	mou(last 3 months)	...	5G Data	Terminal Types	Offsetting Phone Number Users(this month)	Terminal Change(this month)	Residence Areas covered by 5G Tag	Working Areas covered by 5G Tag	label	Balance Arrears_Mean	Data Arrears_Mean	User Data Utilization_Mean
0	2689434779712	26231702691	1	46.0	3.0	5	1	69.10	4487.97	384.67	...	0.0	1	0	0	0	0	0	0.000000	0.0	0.0
1	2697442927197	27358921188	0	53.0	2.0	1	0	15.03	0.02	272.67	...	0.0	0	0	0	0	0	0	9.026667	0.0	0.0
2	2697596026162	25912868422	1	30.0	2.0	1	2	25.31	1401.64	95.00	...	0.0	1	0	0	0	0	0	0.000000	0.0	0.0
3	2694519728363	25988134864	0	41.0	3.0	1	1	16.45	1303.64	29.67	...	0.0	1	0	0	0	0	0	0.253333	0.0	0.0
4	2697510662772	27958259375	1	31.0	3.0	1	1	48.20	2577.91	130.00	...	0.0	1	0	0	0	0	0	0.000000	0.0	0.0

Table2: Label encoder results

4.5 Correlation factor check and Data standardization

We first used pairplot, which examined several major subtypes of variables. We can see that the diagonal line shows the distribution of the individual attributes, while the off-diagonal line shows the correlation plots between two different attributes. Our analysis of several key categorical variables revealed no significant distributional differences, as well as linear correlations between them.

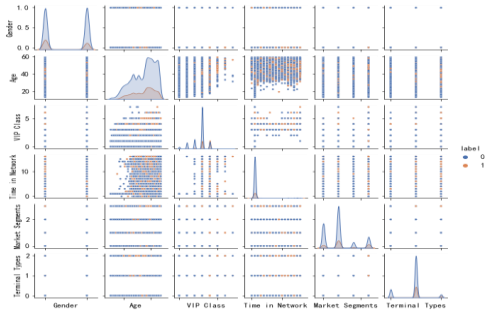

Figure7: Correlation pair plot

We redefined a Pearson correlation coefficient plot to check for the presence of multicollinearity between all variables：

By looking at the correlation coefficients between individual features, the features with strong correlation are recommended to dimension reduce.

We chose one of the more irrelevant variables for each pair of columns with multicollinearity features to reduce the dimensionality of the features, and after consideration, we choose to remove the following columns: Broadband Bandwidth，User Main Package，Average ARPU(last 3 months)，Working Areas covered by 5G Tag.

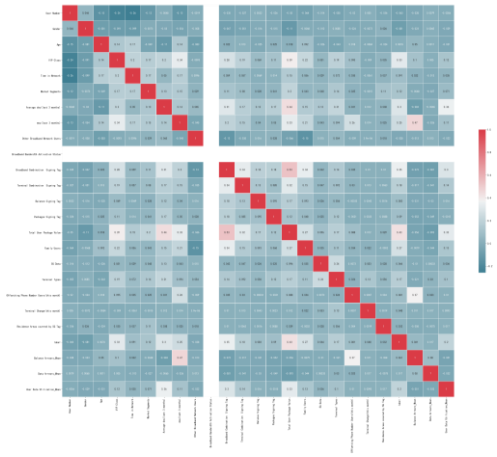

Figure8: Pearson correlation coefficient plot 2

After processing we found from the correlation coefficient plot that the multicollinearity has been eliminated. Then we normalize the data using the following functions: $X_{std} = \frac{X - X_{min}}{X_{max} - X_{min}}$

5. Model building and evaluation

For the latent customer classification problem, we chose to use three algorithms, decision trees, random forests, and logistic regression, to build the model and optimize it by evaluating the parameters.

5.1 Model effect comparison

5.1.1 Logistic Regression

The accuracy of the logistic regression model was around 80%, with 100% recall for non-5G users but 0 recall for 5G users, and the AUC metric was also low, making the model ineffective.

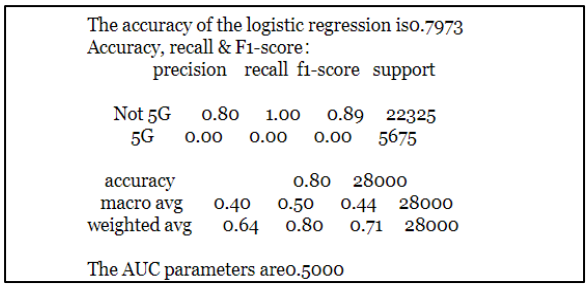

Figure9: logistic regression model performance

5.1.2 Decision Tree Model

The accuracy of the decision tree model was around 81%, with a recall rate of 88% for non-5G users and 54% for 5G users, and an AUC metric of 0.7, with the model meeting its performance targets.

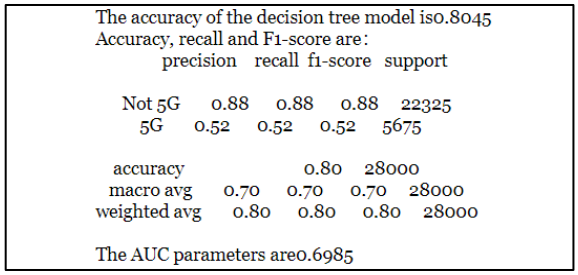

Figure10: decision tree model performance

5.1.3 Random Forest Model

With an accuracy of 86% for the random forest model, a recall rate of 93% for non-5G users and 62% for 5G users, and an AUC metric of 0.77, the model worked optimally compared to the previous two.

Collectively, with default parameters, the random forest model performs best among the three machine learning algorithm models, followed by the decision tree, with logistic regression being the least effective; in addition, the impact of unbalanced sample data may have led to a poor recall rate in predicting 5G users.

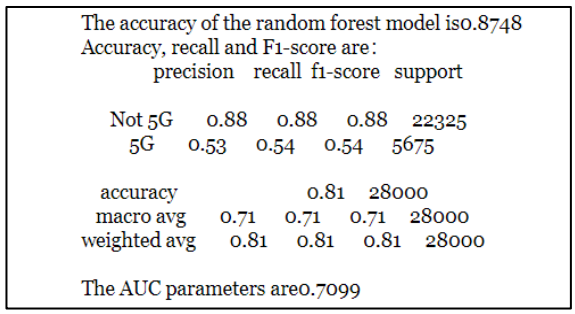

Figure11: random forest model

5.2 Model parameter tuning

Next, parameter tuning is performed for both decision tree and random forest models using a grid search

method.

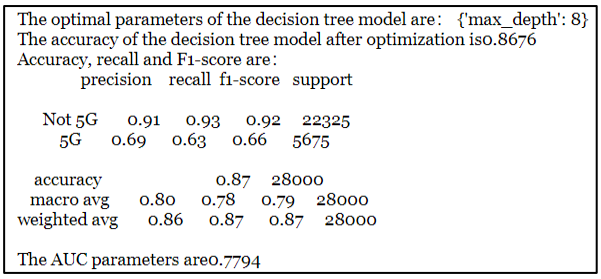

Figure12: decision tree model after tuning

The results of the tuning gave a maximum depth of 8 as the optimal parameter for the tree, an accuracy of 86.76% after tuning, a recall rate of 62% for 5G users and an AUC metric of 0.77, with an improvement in all evaluation metrics.

The tuning results gave a maximum tree depth of 10, a number of trees of 400 as the optimal parameters, an accuracy of 87.03% after tuning, a recall rate of 60% for 5G users, and an AUC metric of 0.77.

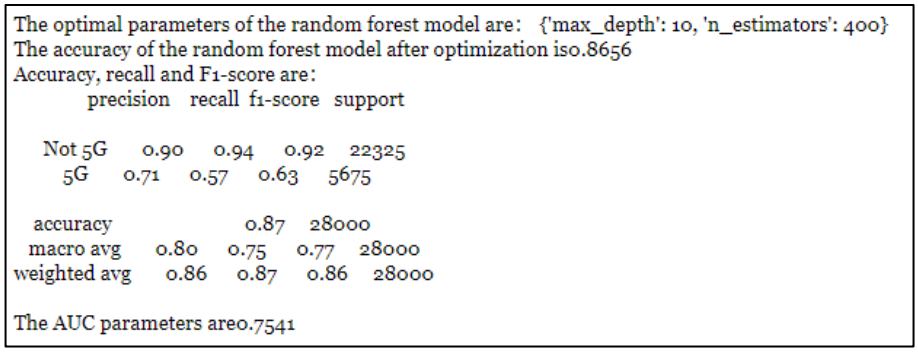

Figure13: random forest model after tuning

In summary, there is little difference in performance between the decision tree and random forest models after adding grid search, and both can be considered.

6. Model deployment and application

We apply our latent guest prediction model to two test sets result_predict_A and result_predict_B and summed the number of customers with an opening probability greater than 0.5 and obtained the following results:

1. There are 10,000 data in the dataset area A. According to the model prediction, 2,948 data have 5G package switching intention, accounting for about 29.48%.
2. There are a total of 50,000 data in the dataset area B, of which 14,650 are predicted by the model to have an intention to switch to a 5G package, accounting for about 29.30%.

For the classified dataset, we can propose some targeted strategies for business analysis of the dataset, primarily for user pulling and secondarily for user retention.

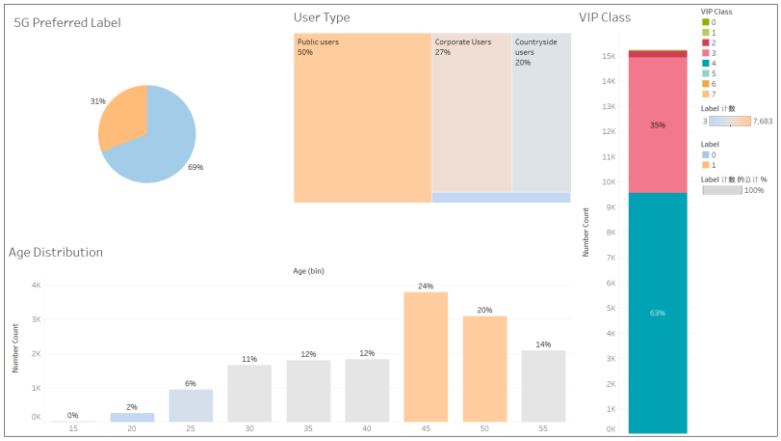

Figure14: Customer information in region I

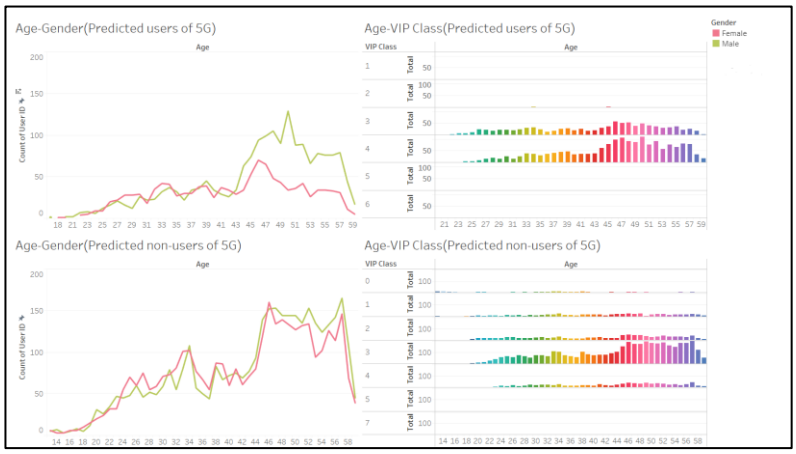

Figure15: Customer information in region 2

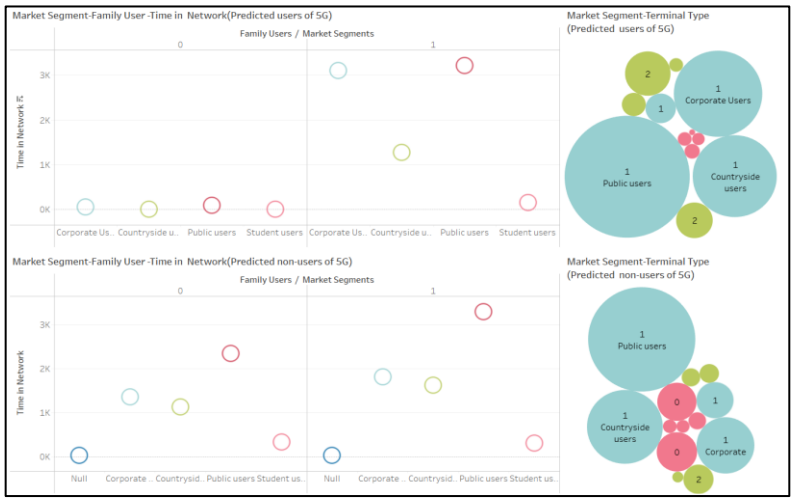

Figure16: Customer information in region 2

The recommended marketing strategy is as follow:

For VIP 4 customers, instead of simply giving discount, put more advertisement on how convenience 5G could be and what problems can 5G solve on their daily life to persuade them to buy 5G,this could be made For VIP 3 customers, try to give some discount and hold more events to attract customers to improve their VIP class, like giving membership point redemption (phone bills, gifts, etc.), or increase the amount of bonus points after opening 5G services.

For public users, use methods ranging from increasing our promotion efforts, such as advertising, sending SMS. For countryside user, our advertisement to them can be publicizing with 5G how easy can family members make video call, cause that's what they most likely to care. And those advertisements can be put on SMS like we used to, we should put them on TV cause that's the main media people in countryside are using.

For operators whose terminal type is 0/2, they should continue to improve 5G functions, enhance mobile broadband and fixed wireless access, realize large-scale machine-type communication and key real-time communication functions, and increase their performance will attract more users Choose to upgrade 5G services. In addition, regarding the current problem that 5G service packages are still more expensive than 4G services, operators should find ways to reduce costs, reduce network operating expenses, and lower the price of 5G packages.

Operators should also improve user experience, combining their advantages of low-latency communication and near real-time response, providing different features for different applications and services, and bringing users an excellent user experience in a variety of application types.

7. Conclusion

For Business aspects, by analyzing the predicted results of users who are likely to use 5g services in region a and b, we believe that the reason for the low number of users is mainly due to the fact that 5g services have just been launched and the market share is small, so we have also summarized some macroscopic marketing strategy to expand users:

First, we can introduce contract smart phones to our users. Secondly, we can improve the accuracy of marketing by diversifying package products and providing personalized services. Third, lower the threshold of 5g usage. Fourth, combine packages with intelligent services. Finally, we should always improve 5g service competitiveness. We can conduct market analysis on competing products. Increase advertising and other promotional efforts to increase market share.

For Technical aspects, this project constructs a model to predict whether 4G users can switch to 5G users by using user information from mobile platforms. Comparing the three models of logistic regression, decision tree and random forest, the decision tree and random forest models performed better, with a prediction accuracy of 87% and a recall rate of 62%.

But both decision trees and random forests have certain disadvantages. For example, decision trees, even when repruning is done, it often over-fits and has poor generalization performance. For this purpose we can subsequently introduce some other machine learning models： GBDT, AdaBoost Classifier, Logistic Regression CV, SGDClassifier, LinearSVC,etc.