A Comparative Study on Contract Recommendation Model: Using Macao Mobile Phone Datasets

GANG WANG\textsuperscript{1} AND NAIQI WU\textsuperscript{1,2}, (Fellow, IEEE)

\textsuperscript{1}Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau
\textsuperscript{2}State Key Laboratory of Precision Electronic Manufacturing Technology and Equipment, Guangdong University of Technology, Guangzhou 510006, China

Corresponding author: Gang Wang (gangtt@189.cn)

This work was funded in part by the Science and Technology Development Fund (FDCT), Macau SAR (file Nos. 0017/2019/A1).

\textbf{ABSTRACT} Bordering on the Mainland of China, Macao is a city with diverse population and is also a world-famous gambling city. The residents living in Macao are not only local residents, but a large-scale of residents are from the Mainland of China. There are diversified living habits among different resident groups. Some of them live and work in Macao most of the time, while others cross the border between Macao and the Mainland of China frequently. These diversities result in different demands for mobile phone contract services. Machine learning and data mining are powerful tools used by telecom companies to monitor the behavior of their customers. This paper aims to build a contract service recommendation model suitable for Macao telecom companies, which can accurately recommend the best fit contract services according to call data records (CDR). Based on data mining algorithm and large number of comparative experiments, this study conducts a study on variety of factors that have effect on the contract recommendation for obtaining a composition structure of factors that have the greatest impact on contract recommendation accuracy so as to improve the operation efficiency of the model. In addition, this study makes comparisons among five classification algorithms, including Bayesian, logistic, Random trees, Decision tree (C5.0), and KNN (k-nearest neighbor). The results are analyzed by different metrics such as Gain, Lift, ROI (region of interest), Response, and Profile. Experimental results demonstrate that the best classifier is decision trees (C5.0), and the contract service obtained by this method can achieve a recommendation accuracy close to 90%, which successfully reaches the expected goal.

\textbf{INDEX TERMS} Business intelligence, contract recommendation, data mining, exploratory data analysis.

\section{I. INTRODUCTION}

It is known that the Internet is a technological breakthrough in the 1990s, but it is the emergence of mobile phones that really changes the way we communicate. In 2018, the number of mobile phones in the world exceeded 7.9 billion, including more than three billion smartphones. China is the world’s largest mobile phone market, and also the fastest growing market. The number of mobile phones in The Mainland of China has reached 1.59 billion in 2019, with a penetration rate of more than 110%, equivalent to more than 110 mobile phones per 100 people. In Macao, the number is even more surprising. With a population of 670,000 only, Macao has more than 2.5 million mobile phones and a penetration rate of more than 300%.

For telecom operators, a critical issue that must be considered when designing a communication contract is how to maximize the benefits while being accepted by a customer. At the same time, a customer has different preferences, such as the demand for cross-regional Internet traffic, the demand for specific mobile applications, the demand for international long distance and roaming calls, and the demand for night traffic. Also, the communication behavior of a customer would change over time, which results in that a constant communication contract cannot always meet the customer’s needs. According to customer’s call data records (CDRs), studying customer’s mobile phone usage preferences can help telecom operators to design more humanized communication contracts and achieve a win-win situation between a mobile phone enterprise and its customers [1].

For Macao’ telecom operators, revenue can be improved by recommending customers to handle higher price contracts,
but the key is whether the recommended contracts are appropriate. In the past, marketers did not use scientific and reasonable methods to recommend package contracts. Such a recommendation method depends on human subjective factors, and surely the selected contracts are not necessarily the best. Therefore, this paper proposes a data mining method to help customers choose contracts.

Based on machine learning algorithm, there are many researches on CDRs. Previous studies mainly focus on customer clustering, urban population mobility, and customer churn early warning. Note that contract service recommendation is one of the important applications of machine learning and up to now less attention is paid to this issue. This paper uses the customer data of Macao as the dataset to carry out this study due to the typical research significance of Macao’s diversified population. Studying the differences in communication behaviors between Macao residents and Chinese residents in Macao can help a mobile phone operator to better understand customers’ choice preferences and choice basis for contract services. It is beneficial for telecom companies to recommend contract to customers more accurately and design more popular packages. In short, the main contributions of this paper are as follows:

- According to the location information of CDR, CDR is transformed into local airtime, international direct dial (IDD), roaming, local data traffic, China main land data traffic, Hong Kong data traffic, and short message service (SMS). This data consolidation method is consistent with the design principle of a contract service, which can more accurately reflect the real preferences of customers.
- Using the decision tree model to get the ranking of factors’ main score, and according to this ranking, it determines the best combination of input factors through large number of experiments.
- Five classification algorithms, including Bayesian, logistic, random trees, decision tree (C5.0), and KNN (k-nearest neighbor), are compared and analyzed. Through different evaluation methods, such as Gain, Lift, ROI, Response, and Profile, it is proved that decision tree (C5.0) is the optimal contract recommendation algorithm.

This paper is organized as follows: Section 2 briefly presents the related work on several relevant concepts such as mobile phone datasets construction, city users’ classification, and the machine learning classifiers used in this study. Section 3 describes the experimental setup and discusses the experimental results with comparisons. Section 4 concludes the study.

II. RELATED WORK
A. MOBILE PHONE DATASET CONSTRUCTION
With the rapid development of mobile communication in recent years, the popularity of mobile phones is increasing. The rapid growth of the number of mobile phones has a huge impact on people’s economic and social life. With the development of mobile phones, mobile sensors have been widely used because of their convenience. These sensors record personal conversations, movements, and activities [2]. Due to the wide applications of smart phones, sensor information has become very popular. At present, many researches have been done [3]–[4] for collecting large number of human behavior data sets to further understand human interaction. These data are used to monitor urban transport and human activities [5], [6]. For example, it helps to monitor population movement in urban areas [7]–[9] or understand the spread of diseases in real-time [10]–[12]. The widespread use of mobile applications provides opportunities for financial transactions [13] and entertainment applications [14] through mobile devices.

CDR has four major types [15]: voice call, SMS, MMS (Multimedia Messaging Service), and data traffic. In the past, voice call and SMS were the representative of traditional communication. With the popularity of mobile Internet and smart phones, MMS business has been eliminated, but data traffic has increased, meanwhile voice calls and SMSs remain stable with slight reduction. For a Voice call or SMS record [16], it contains the encrypted cell phone numbers of a caller and callee, timestamp about date and time of the call, duration and the initial cell tower ID [17]. With the latitude and longitude of a local cell tower dataset, the caller and callee’s geographic locations can be recognized. Some operators separate a voice call into two independent records, outbound and inbound calls [18]. The former only contains a caller’s initial location, while the latter contains callee’s location [19]. When both caller and callee are clients of an operator, two locations are provided. For data usage records, in comparison with voice call records, a callee’s phone number is not included, but data usage is added with KB as measurement units. Data usage records have a higher occurrence frequency than voice calls and SMSs, because smartphones keep online all the time, APPs connect to Internet every a few minutes in background, even users do not use cellphones [20]. In the last decade, studies on CDR focus more on voice call records, which is considered as an ideal dataset for social relationship [21] and mobility researches [22]–[24]. However, recently, with the highspeed development of social networks such as Twitter, Facebook, and Weibo, many researchers have published their social relationship work of using the data from social networks instead of voice call records [25]–[27]. Nevertheless, in the human mobility and urban activity research field, CDR with data traffic is still showing great value. In countries where smart phones are popular, data usage records provide an approach to get location data source with low cost and nearly in real-time.

Another kind of valuable information from CDR is the cell phone user individual data, including user name, age, gender, contract plan, registration ID certificate type, billing address, and average revenue per user (ARPU) [28]. However, for the personal privacy reason, most of CDR data used in the
literature are anonymized, but some studies got individual social characteristics information with help of local operators. For example, Frias Martinez proposed a study on gender characteristics and automatic recognition of mobile phone users in developing economies based on behavioral, social and mobile variables, using CDRs of about 10,000 users from developing countries, whose gender is a priori [29], [30].

B. CITY USERS’ CLASSIFICATION

For analyzing user’s calling behavioral tendencies, Macao’s mobile users can be divided into residents, commuters, and visitors, which realizes the convenient reading and reasoning of big data. The data come from the mobile network and record a user’s position in a call. The categories of behavior derived from the definition of the demographic agency are described below. Given reference area A [31]:

- A **Resident** is an individual who lives and works in A, so his/her presence in A is important for all time periods.
- A **Commuter** is an individual who lives in a different zone B but works/studies in zone A. It is expected that the presence in Zone A is almost entirely concentrated on work/study days and work/study hours.
- A **Visitor** is an individual who lives, works/studies outside of A and visit A once or occasionally.

C. CLASSIFICATION METHODS

Supervised learning is applied to classification problems. Its learning process is divided into two stages: training and testing. In the training stage, the training data set is used to construct a classification model with known target class labels. Then, in the test phase, the unknown instances of a target class label are classified by the generated model. Different classification algorithms are given in Table 1 and they are briefly introduced as follows.

1) BAYESIAN CLASSIFICATION

Bayesian classification [32] can predict the probability of class members. The effect of attribute values on a given class is independent of the values of other attributes, which is assumed by naive Bayes algorithm. Naive Bayes algorithm is expanding in the number of prediction and rows, and quickly building the model. The prediction probability is derived from naive Bayesian algorithms. The probability of occurrence of event $X$ of given event $Y$ ($P(X|Y)$) is directly proportional to the probability of occurrence of event $Y$ of given event $X$ multiplied by the probability of occurrence of event $X$ ($P(Y|X)P(X)$) [33]. The Bayesian formula is shown below:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (1)$$

2) RANDOM TREES

Random Trees is a classifier which contains many decision trees. Each node of a decision tree represents a resolution rule of the subset attribute. By voting on trees to produce results, this process minimizes overtraining. A research [34] revealed that random numbers are effective in cost sensitive learning, sampling techniques, and predicting customer churn.

3) K-NEAREST NEIGHBOR (KNN)

A classifier based on KNN instance operates on unknown instances [35]. According to some functions, one can classify unknown and known. In KNN classification, unknown samples are given, and the pattern space of K training samples is searched by the classifier. These samples are closest to the unknown samples. Proximity is defined by Euclidean distance. Unknown samples are assigned to the most common class among its K nearest neighbors. Euclidean distance is shown below:

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \quad (2)$$

4) DECISION TREE

This method uses tree structure to establish classification model. It divides the dataset into smaller subsets. Leaf nodes represent decisions. A decision tree classifies the cases according to their eigenvalues. Each node represents the characteristics to be categorized in the decision tree, and each branch represents a value. Examples are categorized from root nodes and sorted according to their eigenvalues. Classification and numerical data can be processed by decision trees [36].

5) LOGISTIC REGRESSION

Logistic regression is a probability statistical model, and two variables are used to predict categorical variables. The forecast variable depends on one or more variables, either numerical or nominal [37]. A study [38] shows that after data conversion, logistic regression performs well. Assume that there are $m$ samples of the pairs $(x_i, y_i) , i = 1, 2, \ldots, m, y_i \in \{-1, 1\}$ is a binary class label for each sample $i = 1, 2, \ldots, n$. Then, by the logistic regression for binary classification, the occurrence probability of the class is modeled as follows:

$$P(y = \pm 1|x, w) = \frac{1}{1 + \exp \left(-y (w^T x + b) \right)} \quad (3)$$

where $b$ is the intercept, $w = (w_1, w_2, \ldots, w_k)^T$ are parameters to be calculated.

D. C5.0 CLASSIFIER

C5.0 is another new decision tree algorithm developed by Quinlan on the basis of C4.5 [54]. It contains all the functions of C4.5 and integrates a series of new technologies, the most important of which is “boosting” [55] technology to improve the accuracy of sample identification. However, C5.0 using boosting algorithm is still in progress, which is not available in practical application. C5.0 algorithm has many characteristics, such as:

- Large decision trees can be regarded as a set of rules that are easy to understand.
### TABLE 1. Studies in recommendation using classification methods.

| Classification Methods       | Papers                  | Result(s)                                                                 |
|-----------------------------|-------------------------|---------------------------------------------------------------------------|
| Bayesian                    | [39]                    | Propose a novel approach for constructing users’ movie preference models   |
|                             | [40]                    | Develop a method for recommending products to customers with applications to both on-line and surface mail promotional offers |
|                             | [41]                    | Use a hierarchical Bayesian approach to model user preferences in different contexts or settings |
| Random Trees                | [42]                    | Propose a promotion recommendation method and system based on random forest to analyze the customer profiles and historical mobile data usage |
|                             | [43]                    | Propose a system which uses random forest as multilevel strategies to predict recommendations based on users’ ratings while targeting users’ mind-set and current trends |
|                             | [44]                    | Predict the classification of RNA hybridization results                     |
| K-Nearest Neighbour        | [45]                    | Propose a recommendation system which uses k-nearest neighbor algorithm to determine the top n product recommendations for each buyer |
|                             | [46]                    | Identify customer click stream data and recommend customized browsing options to meet specific needs on RSS readers |
|                             | [47]                    | Proposes a learning community recommendation approach to facilitate that a learner joins the appropriate learning communities, |
| Decision Tree              | [48]                    | Find the customers with the greatest loss probability and recommend promotions |
|                             | [49]                    | Identify the main characteristics of consumers in each segment             |
|                             | [50]                    | Improve the recommendation accuracy of value-added service (VAS) and wireless data service |
| Logistic regression        | [51]                    | Achieve higher song recommendation accuracy rate and recall rate in a music website |
|                             | [52]                    | Established an e-commerce user churn prediction model on e-commerce customer churn behavior |
|                             | [53]                    | Predict the polarity (positive or negative) of a given movie review        |

- C5.0 algorithm confirms noise and lost data.
- C5.0 algorithm is used to solve the problem of over fitting and error pruning.
- In classification technology, C5.0 classifier can predict which attributes are related and which attributes are independent of classification.

### E. INFORMATION GAIN

A decision tree is a collection of branches, leaves (indicating the quality of credit rating), and nodes (specifying the tests to be performed). Decision trees and rules classify observations identified by variables. They divide a region of variable space into sub regions recursively according to the variable with the largest amount of information. In order to determine the criterion of selecting the variable with the largest amount of information, the information gain ratio criterion is adopted [56]. Other methods for selecting variables include v-square contingency table statistics, Gini coefficient or g-statistics. The information gain criterion is based on Shannon’s information theory, which points out that the information of an event is inversely proportional to its probability, and can be measured in bits by subtracting the base 2 from the logarithm of the probability. The observed value belongs to group \( C_i \) (i.e., good or bad credits) with an a priori probability. The information (number of bits) required to identify two sets of observations \( I \) is:

\[
I(S) = - \sum_{i=1}^{2} P(C_i) \log_2 P(C_i) \tag{4}
\]

where \( I(S) \) = entropy of information of data set \( S \), \( P(C_i) = a \) priori probability for an observation that belongs to group \( C_i \).

Next, a test is performed on Variable \( A \) to create branches and partition data set \( S \) into \( n \) subdivisions. The expected information requirement after partitioning is then:

\[
I(S, A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times I(S_i) \tag{5}
\]

By branching variables, the two groups of good credit and bad credit need less information; this difference is called information gain or mutual information:

\[
Gain(A) = I(S) - I(S, A) \tag{6}
\]

In order to establish decision tree effectively, it is very important to branch the variables that get the most information. If one uses the above measures to select the variables with the largest amount of information, one tends to test partitions in multiple sub partitions. Therefore, the normalized gain criterion is normally preferred:

\[
Gain ratio(A) = \frac{Gain(A)}{I(A)} \tag{7}
\]

The induction of a decision tree is a process of maximizing information entropy by selecting variables as branches, and it can also be regarded as the process of maximizing information gain. When a node of a tree contains observation values in the same group only, the entropy is equal to zero, which means that the classification decision is defined for the observation values belonging to the node [57]. This process is repeated until both groups are fully defined. A variable can be used multiple times in a tree.

### F. SPSS MODELER

IBM SPSS Modeler is a data mining and text analysis software application developed by IBM. It is used to build prediction models and perform analysis tasks. Its visual interface allows users to utilize statistical and data mining algorithms without programming.

The cross-industry process for data mining (CRISP-DM) methodology is a general modeling process, which can be applied to various industrial and commercial problems. According to CRISP-DM, we propose the following five modeling steps.
Five-Phase Modelling Cycle From CRISP-DM Perspective:

- Problem understanding, including problem objectives, assessment, data mining objectives, and project planning. This is the most important stage of data mining.
- Data understanding, including the collection of raw data, data understanding, data exploration and verification of data quality. Data preparation, including selection, cleaning, construction, integration and discretization of data.
- Modeling. This phase includes selecting modeling techniques, generating test designs, establishing and evaluating models.
- Assessment. Evaluate how data mining results help an analyst achieve their goals. This phase includes evaluating the results, reviewing the data mining process, and determining the next step.
- Deployment. Focus on integrating new knowledge into daily business processes to solve initial business problems. This phase includes planning deployment, monitoring and maintenance, generating final reports, and reviewing projects.

III. EXPERIMENTAL STUDY

This part uses the user data of Macao telecom operator to carry out the modeling process based on the decision tree model and evaluate the model.

A. MODELLING APPROACH

The modeling method is shown in Figure 1. It begins with exploratory data analysis, summarizes data visually, finds patterns and data exceptions, preprocesses data, formats incomplete, inconsistent and/or missing data, and then performs feature engineering to select features that most likely affect the attributes of the target class.

1) EXPLORATORY DATA

The single variable frequency analysis is used to describe the exploratory data analysis (EDA) and describe the key characteristics of each attribute, including minimum and maximum values, average values, standard deviation, etc. It is also used to generate value distributions and identify missing and outliers.

α: DATA SELECTION AND IMPORT

In this experiment, the data mining source includes two parts: customer information and call data records. In particularly, ARPU and CDRs in Table 2 are the average values of three months.

β: DATA INTEGRATION

Data integration refers to the combination of data from multiple data sources to form a complete data collection, which may include multiple databases or files. Data integration is not a simple data consolidation, but a unified and standardized processing of heterogeneous data. In this paper, we combine CDRs to facilitate data analysis. By data integration, we get Table 3 from Table 2.

c: DATA UNDERSTANDING

First, this study compares the ARPU characteristics of different age groups, as shown in Figure 2. It shows that the users with age being 45-59 are the most part of the users,
reflecting that the overall users are older in this batch of experimental data. In the ARPU distribution, the users of ARPU MOP100 to MOP150 are the most, and more than half of the customers have ARPU less than MOP150.

Besides, this study compares the differences in communication habits between Macao residents and the residents of Mainland China. Macao residents belong to local residents, working and living in Macao, while the residents of Mainland China belong to commuter, working/studying in Macao and most of them do not live in Macao, but live in Zhuhai, a neighbor city of Macao in Mainland China. The residents of Mainland China often travel between Macao and Mainland China, their demands for mobile data and roaming calls in Mainland China are much stronger than Macao residents. As it can be seen from Fig. 3, in terms of data traffic and roaming call minutes in Mainland China, the demands for the residents of Mainland China are twice as many as that for Macao residents, while IDD and local air time ARPU are almost the same.

The reason for the above phenomenon is that telecom companies provide two different types of contracts: one is Macao local contract, which only contains local air time and Macao local mobile data, and in the other contract, the voice and data consumption can be used in both Macao and Mainland China. Each contract type is divided into four grades, as shown in Table 4. The larger the value is, the higher the gear is and the higher the corresponding cost is.

### TABLE 3. New feature after integration.

| Data Source | S# | Feature | New Feature after Integration |
|-------------|----|---------|------------------------------|
| Call Data Record | 7 | Local caller duration | Local airtime |
| | 8 | Local callee duration | IDD |
| | 9 | Mainland China IDD caller duration | Roaming |
| | 10 | HongKong IDD caller duration | SMS |
| | 11 | International IDD caller duration | |
| | 12 | Mainland China roaming caller duration | |
| | 13 | Mainland China roaming callee duration | |
| | 14 | HongKong roaming callee duration | |
| | 15 | HongKong roaming callee duration | |
| | 16 | International roaming callee duration | |
| | 17 | International roaming callee duration | |
| | 18 | Local SMS number | |
| | 19 | Mainland China SMS number | |
| | 20 | HongKong SMS number | |
| | 21 | International SMS number | |
| | 22 | Local Data traffic | Local Data Traffic |
| | 23 | Mainland China Data traffic | Mainland China Data Traffic |
| | 24 | HongKong Data traffic | HongKong Data Traffic |

![FIGURE 3. Comparison of Macao residents and the residents of Mainland China.](image)

2) DATA PREPROCESSING

Data sets can contain noise, missing values, and inconsistent data. Therefore, data preprocessing is very important to improve data quality and time efficiency. It is not advisable to use the original data to build a prediction model, because it would produce weak results. In addition, some machine learning algorithms are not mature enough to extract meaningful information. The data preprocessing technology in this study includes data cleaning, integration, discretization, and reduction.

a: DATA CLEANING

This experiment uses the dataset from Macao Telecom Operator which removes the examples with missing values. The dataset contains 18000 mobile contract users.

b: DATA TRANSFORMATION

Data transformation is a process of normalizing and aggregating data to further improve the efficiency and accuracy of data mining. The nominal to binomial operator is used to convert and map the selected nominal value to its equivalent binomial value. The transformation information from

### TABLE 4. Services included in the contract.

| Services | Contract Type |
|----------|--------------|
|          | Local contract | Two-place contract |
| Macao local call minutes | ✔ | ✔ |
| Mainland China roaming call minutes | × | ✔ |
| Macao local mobile data | ✔ | ✔ |
| Mainland China mobile data | × | ✔ |
| Mainland China IDD minutes | ✔ | × |
nominal to binominal includes gender, ID type, and contract type. Table 5 shows the value range and description of data after transformation.

c: DATA REDUCTION
Data reduction is a process of reducing data representation while still producing similar results. Discretization (or binning) is applied to numeric attributes by converting and grouping consecutive values into discrete categories (or interval levels), because some data mining algorithms accept nominal attributes only and cannot process numeric attributes. Similarly, using original numerical data in the model may have an adverse impact on the performance of the model. For example, ARPU, age, and net days of use in a dataset can be discretized. But, in the experiment, it is found that discretization has little effect on the calculation speed, so this experiment skips the discretization process.

3) FEATURE ENGINEERING
Feature engineering allows the transformation of raw data into features to help improve the overall performance of the prediction model. In most cases, in order to improve the accuracy of the algorithm, it is necessary to reduce the number of features. Features that may have an impact on the target attributes are retained, while features that may be overly suitable for the model are discarded. Feature selection is applied to reduce the number of predictors. By doing so, it reduces the number of errors and improves the accuracy. The data set is divided into training samples and test samples. The mining model is established by using the training set, and the correctness of the model is verified by the test set.

4) MODEL BUILDING
In this study, we use the automatic modeling function in SPSS Modeler to provide a fast modeling and verification process. Because of the manual preprocessing before modeling, including replacement of missing values, data conversion, etc., model validation is directly carried out.

SPSS Modeler software provides the relevant features which are important to the target attributes. The weights are computed using the Pearson correlation coefficient as displayed in Figure 4. We can see the rank of important factors that affect contract recommendation.

By combining the content of a contract with the importance ranking of the influencing factors, it is found that the last several factors are of low importance and not considered in the contract. It has little effect on the accuracy of contract recommendation. If these input values are removed, will the accuracy of contract recommendation be affected? With this problem, the following tests are done in this paper:

Test 1. Take contract type as the target value and all other factors as the input value. The accuracy is 86.33%

Test 2. Excluding Hong Kong data traffic, SMS and gender, the accuracy is dropped to 85.2%

Test 3. Excluding ID type and age, the accuracy is increased to 88%

Test 4. Excluding local air time, the accuracy is increased to the highest 88.4%

![FIGURE 4. Rank of important factors affecting contract recommendation.](image-url)
B. MODEL EVALUATION

In this paper, we choose the best input values and the best classification algorithm by comparing different algorithms on contract recommendation accuracy. In addition, the rationality of the algorithms is verified by evaluating the values of Gain, Lift, ROI, Profit, and Response.

The data set is divided into training samples and test samples. 70% of the data is used as training samples to establish the data model, and 30% of the data is used as test samples to verify the correctness of a model. If the prediction results meet the needs of the predetermined business objectives, the prediction model can be applied to actual situations, through existing customer resources of the telecom enterprise with the model prediction, to select the most suitable package for customers, in order to reduce customer churn.

This section compares the performance of different classifiers to select the best method for contract recommendation. In order to select a most suitable classification algorithm for contract recommendation, we compare five classification algorithms: Bayesian, logistic, random trees, C5.0, and KNN (k-nearest neighbor). By comparing these algorithms on the accuracy of contract recommendation, the comparison is done based on five indicators: gain, lift, ROI, profit, and response. In this way, the optimal algorithm of contract recommendation can be determined.

1) **GAIN**

For gain, the ideal situation should be to quickly reach a very high cumulative gain and quickly tend to 100%. As shown in Figure 6, C5.0 is the algorithm that can reach the highest cumulative gain as soon as possible and approach to 100%.

2) **LIFT**

For lift, it is ideal to maintain a high lift value for a period, or slow down for a period, and then quickly drop to 1. As shown in Figure 7, C5.0 is slightly ahead of Random Trees, both of which maintain a long period on the higher cumulative lift, and then rapidly decline to 1. Other algorithms are not very good with respect to C5.0 and Random Trees.

3) **ROI**

For ROI, the ideal situation is to maintain a segment on the higher cumulative ROI and then rapidly drop to the general level. Similar to lift’s results, from Figure 8, we can see that C5.0 and Random Trees are in the lead, and the other algorithms are not very good.

4) **RESPONSE**

For response, the ideal situation is to maintain a period of high cumulative response, and then decline rapidly. From Figure 9, it can be seen that C5.0 and Random Trees are in the lead.

5) **PROFIT**

The ideal situation is that it should rise rapidly in the early stage and descend rapidly after the vertical coordinate of 50%.
behavior more accurately so as to recommend the best fit contract to improve customer satisfaction and reduce churn rate.

This study applies exploratory data analysis and feature engineering on a telecom dataset and use these techniques to improve the performance of five classifiers for contract recommendation. This study also discusses the prediction results of different classifiers. Experimental results demonstrate that C5.0 outperforms the other classifiers in almost all evaluation metrics. Finally, this study recommends C5.0 as an algorithm for contract recommendation modeling.

At the same time, there are still some problems to be studied. First, does ARPU really reflect customers’ consumption level? In real life, we find that, for a kind of customers, there is a phenomenon called “consumption inhibition”. For example, for this kind of customers, when the data traffic of a contract for them is going to be exhausted, they then try to reduce the mobile Internet behavior in the remaining days of the month, so as to avoid extra consumption beyond the contract. However, such customers do not resist spending more money to upgrade the contract service but are reluctant to exceed the cost of the contract. How to calculate the real data consumption demands and acceptable price of such customers? Second, if a customer has multiple phone numbers in different telecom operators, how to calculate the real ARPU of such a customer? Third, in the actual contract recommendation process, apart from the price and contract service factors, are there other important factors that affect the selection of contract service, such as signing a more expensive contract due to a new mobile phone? Due to the use of an APP, more data is needed, so does the idea of upgrading the contract arise?

In the future, we plan to conduct research on the behavior phenomenon of “consumption inhibition” and study an algorithm that can calculate the potential consumption level of customers. It also focuses on the correlation between APPs and mobile data usage, clustering and promotion recommendation based on the content data of mobile Internet.

**REFERENCES**

[1] C. Peera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Sensing as a service model for smart cities supported by Internet of Things,” Trans. Emerg. Telecommun. Technol., vol. 25, no. 1, pp. 81–93, Sep. 2013.

[2] N. Győrbíró, Á. Fábián, and G. Hományi, “An activity recognition system for mobile phones,” Mobile Netw. Appl., vol. 14, no. 1, pp. 82–91, Nov. 2008.

[3] M. S. Hashemian, K. G. Stanley, D. L. Knowles, J. Calver, and N. D. Osgood, “Human network data collection in the wild: The epidemiological utility of micro-contact and location data,” in Proc. 2nd ACM SIGHIT Symp. Int. health Informat. (IHI), Miami, FL, USA, 2012, pp. 255–264.

[4] S. M. V. Gwynne, “Improving the collection and use of human egress data,” Fire Technol., vol. 49, no. 1, pp. 83–99, Jan. 2011.

[5] R. Ahas, A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook, “Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn,” Int. J. Geographical Inf. Sci., vol. 29, no. 11, pp. 2017–2039, Jul. 2015.

[6] W. Tu, J. Cao, Y. Yue, S.-L. Shaw, M. Zhou, Z. Wang, X. Chang, Y. Xu, and Q. Li, “Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns,” Int. J. Geographical Inf. Sci., vol. 31, no. 12, pp. 2331–2358, Jul. 2017.

---

**TABLE 7. Ranking of accuracy.**

| Rank | Classification Methods | Accuracy |
|------|------------------------|----------|
| No.1 | C5.0                   | 88.4%    |
| No.2 | Random Trees           | 79.7%    |
| No.3 | KNN                    | 78.37%   |
| No.4 | Logistic               | 69.68%   |
| No.5 | Bayesian               | 56.44%   |
[1] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb, “Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti,” PLoS Med., vol. 8, no. 8, Aug 2011, Art. no. e1001083.

[2] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, “Real-time urban monitoring using cell phones: A case study in Rome,” IEEE Trans. Intell. Transp. Syst., vol. 12, no. 1, pp. 141–151, Mar. 2011.

[3] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, “Understanding individual mobility patterns from urban sensing data: A mobile phone trace example,” Transp. Res. C. Emerg. Technol., vol. 26, pp. 301–313, Jan. 2013.

[4] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding the spreading patterns of mobile phone viruses,” Science, vol. 324, no. 5930, pp. 139–142, May 2009.

[5] P. Wang, M. C. González, R. Menezes, and A.-L. Barabási, “Understanding the spread of malicious mobile-phone programs and their damage potential,” Int. J. Inf. Secur., vol. 12, no. 5, pp. 383–392, Jun. 2013.

[6] J. de Monasterio, A. Salles, C. Lang, D. Weinberg, M. Minnoni, M. Travizano, and C. Sarratae, “Analyzing the spread of chagas disease with mobile phone data,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), San Francisco, CA, USA, Aug. 2016, pp. 607–612.

[7] N. Mallat, M. Rossi, and V. K. Tuunnainen, “Mobile banking services,” Commun. ACM, vol. 47, no. 5, pp. 42–46, 2004.

[8] R. Wei, “Motivations for using the mobile phone for mass communications and entertainment,” Telematics Informat., vol. 25, no. 1, pp. 36–46, Feb. 2008.

[9] F. M. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, “Identifying user habits through data mining on call data records,” Eng. Appl. Artif. Intell., vol. 54, pp. 49–61, Sep. 2016.

[10] A. Bose, X. Hu, K. G. Shin, and T. Park, “Behavioral detection of malware on mobile handsets,” in Proc. 6th Int. Conf. Mobile Syst., Appl., Services (Mobisys), Breckenridge, CO, USA, Jun. 2008, pp. 225–238.

[11] W. Jansen and R. Ayers, “Guidelines on cell phone forensics,” NIST Special Publication, vol. 800, no. 101, May 2007.

[12] M. Rod and N. J. Ashill, “The impact of call centre stressors on inbound and outbound call-centre agent burnout,” Manag. Service Quality, Int. J., vol. 23, no. 3, pp. 245–264, May 2013.

[13] V. D. Blondel, A. Decuyper, and G. Krings, “A survey of results on mobile phone interaction patterns based on community structures,” in Proc. IEEE/ACM Int. Conf. Mobile Syst., Apps., Services, and Tech. (MobiSys), Atlanta, GA, USA, Jun. 2011.

[14] L. Gabrielli, B. Furletti, R. Trasarti, F. Giannotti, and D. Pedreschi, “City users’ classification with mobile phone data,” in Proc. IEEE Int. Conf. Big Data (Big Data), Santa Clara, CA, USA, Oct. 2015, pp. 1007–1012.

[15] M. J. Maher, “Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach,” Transp. Res. B. Methodol., vol. 17, no. 6, pp. 435–447, 1983.

[16] A. S. Halibas, A. C. Mathew, I. G. Pillai, J. H. Reazol, E. G. Delvo, and L. B. Reazol, “Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling,” in Proc. 4th MEC Int. Conf. Big Data Smart City (ICBDSC), Muscat, Oman, Mar. 2019, pp. 1074–1079.

[17] P. Cunningham and S. J. Delany, “A review and analysis of churn prediction methods for customer retention in telecom industries,” in Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Jan. 2017, pp. 1–7.

[18] C. S. D. Prasetya, “Sistem rekomendasi pada E-Commerce menggunakan K-Nearest neighbor (KNN) classification method,” in Proc. 13th Int. Conf. Service Syst. Service Manage. (ICSSSM), Kunming, China, Jun. 2016, pp. 1–6.

[19] D. A. Adeniyi, Z. Wei, and Y. Yongguang, “Automated Web usage data mining and recommendation system using K-Nearest neighbor,” Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 4, no. 3, pp. 194–200, Sep. 2017.

[20] D. A. Adeniyi, Z. Wei, and Y. Yongguang, “Automated Web usage data mining and recommendation system using K-Nearest neighbor (KNN) classification method,” Appl. Comput. Inform., vol. 12, no. 1, pp. 90–108, Jan. 2016.

[21] J. W. Li, Y. T. Wang, and Y. C. Chang, “A learning-community recommendation approach for Web-based cooperative learning,” World Acad. Sci., Eng. Technol. vol. 7, no. 3, pp. 19–24, 2013.

[22] Y.-F. Wang, D.-A. Chiang, M.-H. Hsu, C.-J. Lin, and I.-L. Lin, “A recommender system to avoid customer churn: A case study,” Expert Syst. Appl., vol. 36, no. 4, pp. 8071–8075, May 2009.

[23] D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of RPM model-based customer segmentation using data mining,” J. Database Mark. Customer Strategy Manage., vol. 19, no. 3, pp. 197–208, Aug. 2012.

[24] L. Yuan and H. Shu, “Integration of data mining techniques to evaluate promotion for mobile customers’ data traffic in data plan,” in Proc. 13th Int. Conf. Service Syst. Service Manage. (ICSSSM), Kunming, China, Jun. 2016, pp. 1–6.

[25] S. Chang, W. Wei, and X. Guiyuan, “Hybrid recommendation algorithm based on logistic regression refinement sorting model,” in Proc. Int. Conf. Image Video Process., Artif. Intell., Nov. 2019, Art. no. 113212.
[52] Q. Yanfang and L. Chen, “Research on E-commerce user churn prediction based on logistic regression,” in Proc. IEEE 2nd Int. Conf. Netw., Electron. Autom. Control Conf. (ITNEC), Dec. 2017, pp. 87–91.

[53] H. S. Priyanka, B. V. Ramya, and K. Ashok, “Classification model to determine the polarity of movie review using logistic regression,” Int. J. Comput. Sci., vol. 6, pp. 87–91, Dec. 2019.

[54] J. R. Quinlan, “C5.0 data mining tool,” RuleQuest Res., vol. 63, p. 64, Apr. 1997.

[55] E. Alfaro, M. Gámez, and N. García, “Adabag: AnRPackage for classification with boosting and bagging,” J. Stat. Softw., vol. 54, no. 2, pp. 1–35, 2013.

[56] J. T. Kent, “Information gain and a general measure of correlation,” Biometrika, vol. 70, no. 1, pp. 163–173, 1983.

[57] G. Rizk, A. D. Chouakria, and C. Amblard, “Temporal decision trees,” in Proc. 56th Session ISI (Int. Stat. Inst.), Lisbon, Portugal, Aug. 2007, pp. 1–44.

[58] Q. Kang, X. Chen, S. Li, and M. Zhou, “A noise-filtered under-sampling scheme for imbalanced classification,” IEEE Trans. Cybern., vol. 47, no. 12, pp. 4263–4274, Dec. 2017.

[59] Q. Kang, L. Shi, M. Zhou, X. Wang, Q. Wu, and Z. Wei, “A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 9, pp. 4152–4165, Sep. 2018.

[60] W. Tang, Z. Ding, and M. Zhou, “A spammer identification method for class imbalanced weibo datasets,” IEEE Access, vol. 7, pp. 29193–29201, 2019.

[61] H. Liu, M. Zhou, and Q. Liu, “An embedded feature selection method for imbalanced data classification,” IEEE/CAA J. Automatica Sinica, vol. 6, no. 3, pp. 703–715, May 2019.

GANG WANG received the B.S. degree in information security and the M.S. degree in electronics and communications engineering both from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2013, respectively. He is currently pursuing his Ph.D. degree in computer technology and application at the Macau University of Science and Technology, Taipa, Macau. His research interests include data mining and applications, consumer behavior, and customer value in the telecom environment.

NAIQI WU (Fellow, IEEE) received the B.S. degree in electrical engineering from the Anhui University of Technology, Huainan, China, in 1982, the M.S. and Ph.D. degrees in systems engineering from Xi’an Jiaotong University, Xi’an, China, in 1985 and 1988, respectively. From 1988 to 1995, he was with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, and also with Shantou University, Shantou, China, from 1995 to 1998. He moved to the Guangdong University of Technology, Guangzhou, China, in 1998. He joined the Macau University of Science and Technology, Taipa, Macau, in 2013. He was a Visiting Professor at Arizona State University, Tempe, USA, in 1999, the New Jersey Institute of Technology, Newark, USA, in 2004, the University of Technology of Troyes, Troyes, France, from 2007 to 2009, and Evry University, Evry, France, from 2010 to 2011. He is currently a Professor at the Institute of Systems Engineering, Macau University of Science and Technology, Taipa. He is the author or coauthor of one book, five book chapters, and more than 170 peer-reviewed journal articles. His research interests include production planning and scheduling, manufacturing system modeling and control, discrete event systems, Petri net theory and applications, intelligent transportation systems, and energy systems. He was an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART C, the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and the Editor-in-Chief of Industrial Engineering Journal, and is also an Associate Editor of Information Sciences.

* * *

VOLUME 8, 2020

39757