UNIVERSITY OF Southampton

MANG6260

Using Big Data for Consultancy

ID: 30867835

Individual Coursework

Word Count: 2490

2020-5-20

CONTENTS

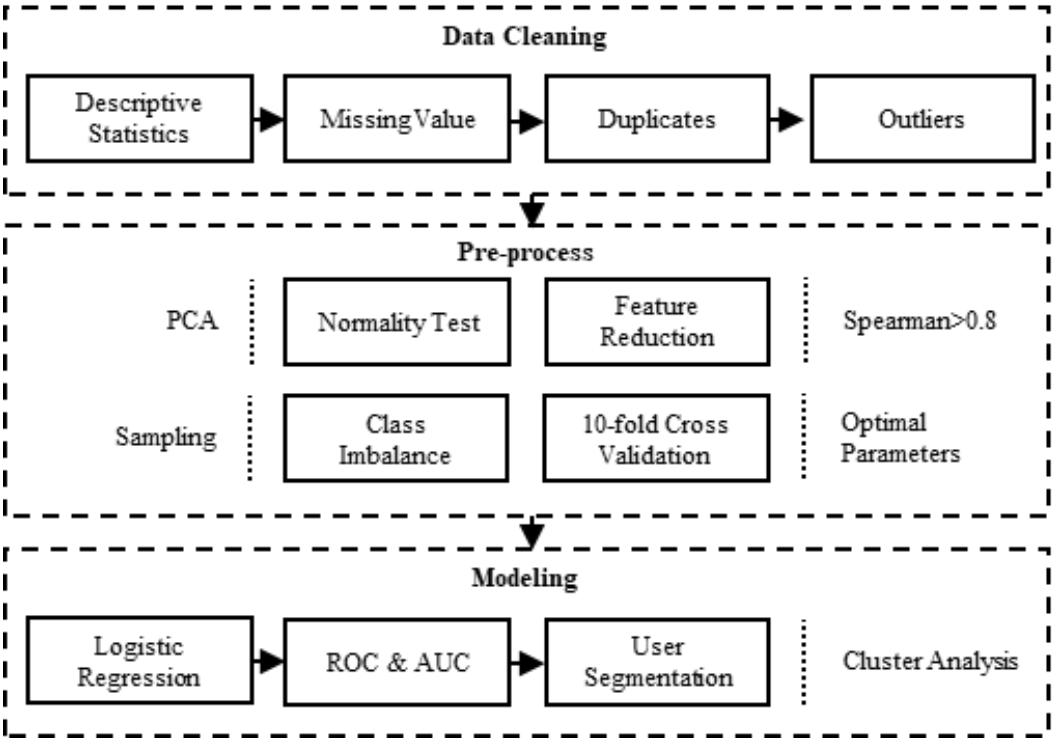
INTRODUCTION

With the rapid development of information and telecom technology, mobile phones have played an indispensable role in people's daily life. Meanwhile, competition in the US telecommunications market is becoming fiercer and gradually saturated. Preserving existing users through churn prediction has become the primary goal of mainstream telecommunications companies (Faris, 2014). As part of customer relationship management methods (Guyon et al., 2014), user churn prediction can effectively help companies reduce customer churn, which is of great significance for the company to increase revenue and improve competitiveness.

Customer churn has been widely discussed, which has led to the advancement of customer churn prediction methods such as logistic regression (Mozer et al., 2000), decision tree (Lima et al., 2009), support vector machine (Archaux et al., 2004), neural network(Hung et al., 2006), triggering algorithm (Aimee et al., 2016), particle classification optimization (Yu et al., 2018) and C5.0 decision tree (Li et al., 2016).

Some studies try to improve the accuracy by ensemble learning. For example, Tsai et al. (2009) proposed an algorithm with high accuracy and stability combined by artificial neural network (ANN) & self-organizing map (SOM algorithm); however, requires extremely high CPU performance. Dalvi et al. (2016) proposed a prediction technique using decision trees and logistic regression. This method can improve the prediction accuracy not much and can only be applied to the problem of less classification. In this report, after data cleaning and random sampling to solve sample imbalance, AUC value obtained by cross validation and logical regression reached 0.7859. Finally, using cluster analysis to subdivide users and make different marketing strategies.

A. ROADMAP



B. PRE-PROCESS

First, this report imports three data into SAS and uses `merge` function to merge them into one data set by `Customer_ID`. Then make descriptive statistics. Do bar chart and pie chart for continuous and categorical variables respectively. Two of them (`tweedie_adjusted` & `handset`) are shown in ***Figure 1***. Then this report deletes duplicates based on `Customer_ID`, a total of 0 duplicates are deleted.

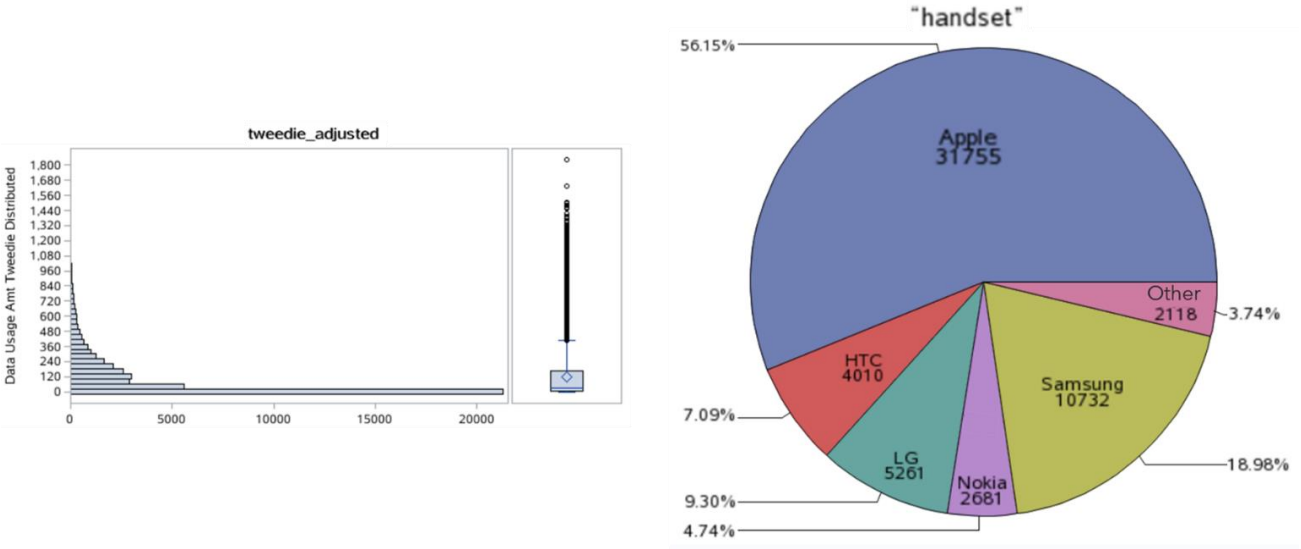


Figure 1 *Descriptive Statistics*

I. Missing Value & Outliers

This report searched the database for missing values and found that variables starting with 'mou_' were particularly missing; therefore, deleted these six variables (`mou_total_pct_MOM` `mou_onnet_pct_MOM` `mou_roam_pct_MOM` `mou_onnet_6m_normal` `mou_roam_6m_normal` `call_category_2`) and created a new data set containing these six variables as back up.

After deleting missing values, 46076 cases remain in this data set.

Box charts and descriptive statistics found outliers in some features. After analysis, this report believes that the outliers are not due to input errors; however, some extreme outliers will become influential points and affect the subsequent model fitting. For ensuring that the data characteristics are not destroyed, this report just deletes three of these outliers.

II. PCA & Normality Test

If the sample covariance matrix is a diagonal matrix, that is, the components of the p-dimensional vector are not related, multivariate normality test can be transformed into p unary normality tests. However, variables are related, and directly diversifying into a unary test will produce errors due to the correlation of the variables. This comes to the principal component analysis (PCA) algorithm, which is a commonly used linear dimensionality reduction method.

This method assumes that all eigenvalues of the sample covariance matrix are greater than 0, and then calculates P independent principal components Zi. The Zi score calculated from the original sample data is used as sample data of p uncorrelated comprehensive variables. At this time, the multivariate normality test has been transformed into p unary normality tests.

In practical applications, a large amount of information of observation data can be provided by the first few principal components, so this report uses the first 2 principal components for normality test results as shown in ***Table 1***.

Table 1 *Normality Tests*

Z1, Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.088272	Pr > D	<0.0100
Cramer-von Mises	W-Sq	153.9044	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	965.3289	Pr > A-Sq	<0.0050

Z2, Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.118677	Pr > D	<0.0100
Cramer-von Mises	W-Sq	184.4376	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1009.114	Pr > A-Sq	<0.0050

In the normality test, the p-values of the D, W-Sq, and A-Sq statistics of z1 and z2 are all less than 0.05. Therefore, it can be considered that the multivariate population does NOT obey the normal distribution. This report attempts to standardize the data and repeat the above test; unfortunately, the multivariate population still does not obey normal distribution.

Since the data after dimensionality reduction through principal component analysis loses its practical significance, this report only uses PCA to test multivariate normality rather than dimensionality reduction.

III. Dimensionality Reduction

Convert 7 character variables (`credit_class sales_channel region handset_age_grp handset lifestage rp_pooled_ind`) into numeric variables to prepare for subsequent modeling.

Delete the following variables: `mfg_samsung mfg_nokia mfg_motorola mfg_lg mfg_htc mfg_apple`. Because their information is given in the `handle` variable.

Because it has been tested before that the multivariate population does not obey the normal distribution, this report deletes variables with strong correlation based on the threshold of Spearman correlation> 0.8. The deleted variables are as follows:

`region_long state_long city_long zip_long forecast_region cs_ttl_hhlds cs_ttl_rural cs_ttl_female res_calls_3mavg_acct data_usage_amt mb_data_usg_roamm01 mb_data_usg_roamm02 mb_data_usg_roamm03 calls_in_pk calls_in_offpk calls_out_pk voice_tot_bill_mou_curr res_calls_6mavg_acct`

IV. Class Imbalance

Machine learning algorithm treats all samples equally, resulting in a high classifier classification accuracy in most classes and low in minority classes. The algorithm has a loss function to be optimized. Taking the binary classifier logistic regression as an example, the loss function is shown in ***Formula 1***. The objective of logistic regression is to optimize the overall accuracy. The errors generated by two classes of misclassification are the same. Consider a $ 23: 1 $ (2354/56557) data set. Even if all samples are predicted to be the majority, the accuracy can also reach 95.8%. Obviously, this is not a good learning effect, so the algorithm has limitations in unbalanced data sets.

This report adopts stratified sampling, and selects 2000 cases at Churn = 1 and 0 respectively. A total of 4000 cases. On this basis, delete character variables that have not been converted into numeric types (because of too many types) and other useless variables. The deleted variables are as follows: `rand Customer_ID upsell_xsell call_center issue_level1 issue_level2 call_category resolution state city product_plan_desc call_category_1 SelectionProb SamplingWeight product_plan_desc`

Next, this report uses **`proc contents`** to check the data type to ensure that all data has been converted to numeric format. For the convenience of coding, this report renames the dependent variable as Y, and the other 74 independent variables are named V2-V75 respectively.

C. MODELING

I. Cross Validation

Cross validation is a practical method for statistically cutting data samples into smaller subsets. In one dataset, most samples are used to train the model, and a small number of samples are predicted by the newly established model. Calculate the sum of their squares through the forecast errors of these small samples. This process continues until all samples have been predicted once and only once. The sum of the squared forecast errors of each sample is called PRESS (predicted Error Sum of Squares).

This report uses 10-fold cross-validation method to calculate the prediction accuracy rate on the test set, that is, to generate 10 examples of cross-validation. In the generated data set of cross-validation, selected = 1 for the behavior training set sample, and 0 for the test set sample. Assign a value to new_y. If selected = 1, Y has value. If selected = 0, the new_y of the row is empty. Next, give the predicted value of new_y is blank.

Set phat as the probability that logistic Regression calculated for each observer belongs to the group. If phat> 0.5, it belongs to the group (here level is 1), otherwise it belongs to another group.

Then this report calculates the prediction accuracy rate (the probability that the accurately predicted samples in the test set for the total predicted samples) and adds cross-validated C statistics to the results to measure the consistency between the observed and predicted values. Finally, the statistical results of the optimal model results are combined to save in the cvparam data set. The cross-validation accuracy of each group is shown in ***Table 2.***

Table 2 *Cross Validation Summary*

	Obs	Replicate	cValue2	acc
	1	1	0.970	0.8950
	2	2	0.970	0.8825
	3	3	0.969	0.8925
	4	4	0.969	0.9050
	5	5	0.970	0.9175
	6	6	0.968	0.9225
	7	7	0.967	0.9225
	8	8	0.967	0.9225
*	9	9	0.967	0.9325
	10	10	0.969	0.9150

* Optimal model selection

II. Logistic Regression

Logistic regression is a classic algorithm for handling binary classification problems. This algorithm forms a linear combination according to the data x and the model parameter θ, then maps the result of

this linear combination to the (0, 1) interval through the Sigmoid function as the classification probability. That is

$$\boldsymbol{P(churn) = h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}}$$

According to the probability threshold phat = 0.5, $\boldsymbol{h_\theta(x)}$ as the classification probability of the data sample z, the classification label of the data sample is obtained. For reducing the risk of overfitting of the logistic regression model, regularization term is added to the cost function. As shown in ***Formula 1***

$$\boldsymbol{J(\theta) = -\frac{1}{m}(\sum_{i=1}^{m}(y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - x^{(i)})))) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2} \qquad \textbf{\textit{Formula 1}}$$

The optimal samples are selected from the 10 data sets of cv with the optimal combination. Take the selected = 1 row corresponding to the optimal group number as the training set, and the rest as the test set to establish the logistic regression model. Besides, this report define the reliability of the P value, which is 90%, and the default is 95%; the modeling method is the stepwise elimination method; the significance of the variable in the model is 0.1, the default is 0.05.

III. Robustness

Because the dependent variable `churn` obeyed the binomial distribution, for the robustness of the model, six independent variables with strong linear correlation with the dependent variable were deleted in this report. The Pearson coefficient correlation of these independent variables and churn are as follows: `count_of_suspensions_6m` (.49), `tot_drpd_pr1` (.33), `nbr_contacts` (.34), `calls_care_acct` (.75), `last_rep_sat_score` (-.34), `price_mention` (. 74).

After these six variables were deleted, 68 independent variables remained. Among them, the dependent variable has the most linear relevant with churn was `network_mention`, and its Pearson coefficient was 0.098.

IV. Model Evaluation

Confusion matrix is a basic tool to evaluate the credibility of the classifier, as shown in ***Table 3***

Table 3 *Basic Confusion matrix*

	Predicted No	Predicted Yes
Actual No	True Positive (TP)	False Negative (FN)
Actual Yes	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensiticity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

In the last step, only the prediction probability is given in the prediction result of the training set, then the observations are classified into specific classes according to the 0.5 boundary. This report adds a column of "pred" (prediction group) to the data set. **`proc freq`** gets the confusion matrix. The confusion matrix of the training set and testing set are shown in ***Table 4.***

Table 4 *Confusion matrix*

Testing Dataset

	Y * pred			
		pred		
	Y	0	1	Total
Frequency Percentage Row percentage Column percentage	0	1284 35.67 71.49 71.45	512 14.22 28.51 28.40	1796 49.89
	1	513 14.25 28.44 28.55	1291 35.86 71.56 71.60	1804 50.11
	Total	1797 49.92	1803 50.08	3600 100.00

Training Dataset

	F_Y * I_Y		
	I_Y(: Y)		
F_Y(: Y)	0	1	Total
0	153 38.25 75.00 74.27	51 12.75 25.00 26.29	204 51.00
1	53 13.25 27.04 25.73	143 35.75 72.96 73.71	196 49.00
Total	206 51.50	194 48.50	400 100.00

AUC is the area covered by the ROC curve (Fawcett, 2006). Its physical significance is: when any pair of (positive and negative) samples are taken, the probability value of the score of the positive sample is greater than the negative, where the score is the confidence that the output belongs to the positive category.

$$AUC = \frac{Sensiticity + Specificity}{2}$$

This report gets sensitivity and specificity obtained through score statement, and then draws the ROC curve. The ROC curve of the training set and the testing set are shown in ***Figure 2.***

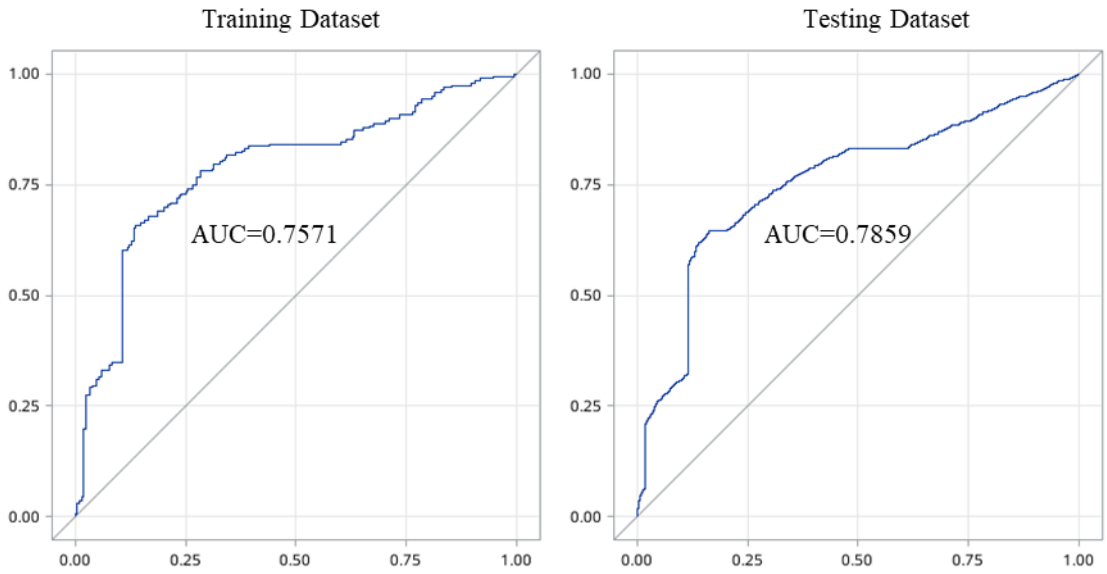
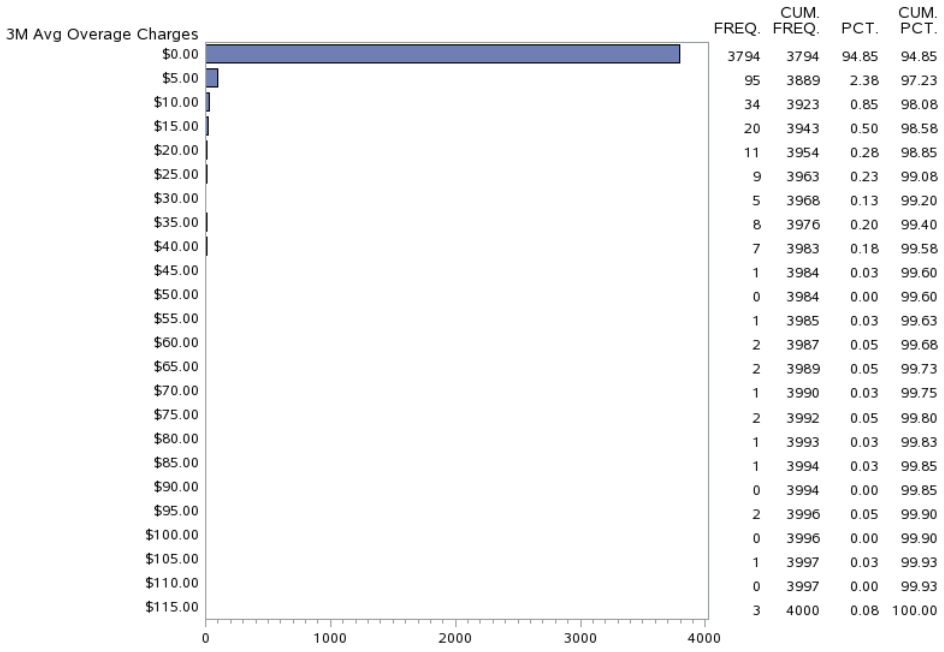


Figure 2 *ROC Curve*

V. Cluster Analysis

This report does consumer segmentation by using cluster analysis on the variable (avg_overage_chrgs_3m). Because the final goal of a company is to earn money. The data set used in this step is the previously balanced data set (2000 cases at Churn = 1 and 0 respectively, a total of 4000 cases). The descriptive statistics of this variable are shown in ***Figure 3***. It shows a typical skewed distribution. Customers who spend $ 0-5 account for 95% of the total.



	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
$0.00	3794	3794	94.85	94.85
$5.00	95	3889	2.38	97.23
$10.00	34	3923	0.85	98.08
$15.00	20	3943	0.50	98.58
$20.00	11	3954	0.28	98.85
$25.00	9	3963	0.23	99.08
$30.00	5	3968	0.13	99.20
$35.00	8	3976	0.20	99.40
$40.00	7	3983	0.18	99.58
$45.00	1	3984	0.03	99.60
$50.00	0	3984	0.00	99.60
$55.00	1	3985	0.03	99.63
$60.00	2	3987	0.05	99.68
$65.00	2	3989	0.05	99.73
$70.00	1	3990	0.03	99.75
$75.00	2	3992	0.05	99.80
$80.00	1	3993	0.03	99.83
$85.00	1	3994	0.03	99.85
$90.00	0	3994	0.00	99.85
$95.00	2	3996	0.05	99.90
$100.00	0	3996	0.00	99.90
$105.00	1	3997	0.03	99.93
$110.00	0	3997	0.00	99.93
$115.00	3	4000	0.08	100.00

Figure 3 *Bar Chart for avg_overage_chrgs_3m*

The selection of the number of clusters is based on R^2 and pseudo F-statistics. The larger R^2 means the smaller the sum of squared deviations in each class, that is, the current number of classifications is

appropriate. However, the more classifications, the larger R^2. Therefore, the number of classifications should be relatively small, meanwhile R^2 should be large enough and no longer increases significantly. The pseudo F-statistics is another indicator to evaluate the effect of classification. If the number of classifications is reasonable, the sum of squared deviations within a class should be small, and the sum of squares between classes should be relatively large. Therefore, the clustering level with a large pseudo F-statistic and a small class number should be taken.

The number of clusters selected in this report is 3, where the pseudo F-statistics = 14423.60, $R^2 = 0.89$.

Result of clusters are shown in the ***Table 5.***

Table 5 *Result of clusters*

	Interval	Quantity
Cluster 1	$ 0- $ 15	3934
Cluster 2	$ 15- $ 59	53
Cluster 3	> $ 59	13
Overall		4000

VI. Limitation

Number of variables. Relying on 74 independent variables, but companies may pay too much for collecting so much data. If inputting few independent variables, the performance of this model may be poor.

Cause of churn. This report only makes predictions about whether customers will churn or not, but it cannot give reasons. That is, only give what, but not why. The cause of customer churn is significant in the operation of the enterprise. If enterprises know reasons of customer churn, it can predict its churn and prevent such churn in a targeted manner. In the future, collaborative algorithms can be used to find reasons.

D. MARKETING STRATEGY

The purpose of data mining is to extract value from data, so the analysis results need to be applied to the actual marketing strategy of enterprises and bring significant profits. Retaining old customers is more important than developing new customers. The cost of developing a new customer is five times greater than retaining an old customer (Kotler, 1994). Customer loyalty is one of the most significant assets that an enterprise can survive in the market. Loyal customers not only reduce the company's service costs (Ganesh et al., 2000), also promote the positive word-of-mouth effect for the company thereby create new transactions for the company.

However, maintaining a customer also requires cost. Companies should first know how much a certain customer churn affects the company. Customers can be ranked according to their contribution value. If A-level accounts for a high proportion of churn customers, it means that the problem is serious. If they are all C-level, then it does not necessarily need to spend huge manpower and financial resources to handle. Companies can identify those A-level churn customers and invite them to participate in in-depth in-person discussions (e.g. Focus group) to understand the reasons for their churn. According to these reasons, design an improvement or recovery plan. Monitor the changes in RFM (Recency, Frequency, Monetary) of all customers, and find people who have a high contribution (high M) but have recently reduced the number of consumptions (R or F). For this group, enterprises can apply some pre-planned discounts. Data mining to predict customer churn and actual marketing strategy before the incident can effectively bring significant benefits to the company.

REFERENCES

Archaux, C., Martin, A., & Khenchaf, A. (2004, April). An SVM based churn detector in prepaid mobile telephony. In *Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications, 2004.* (pp. 459-460). IEEE.

Backiel, A., Baesens, B., & Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, *67*(9), 1135-1145.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1-4). IEEE.

Faris, H., Al-Shboul, B., & Ghatasheh, N. (2014, September). A genetic programming based framework for churn prediction in telecommunication industry. In *International Conference on Computational Collective Intelligence* (pp. 353-362). Springer, Cham.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: an examination of the differences between switchers and stayers. *Journal of marketing*, *64*(3), 65-87.

Guyon, I., Lemaire, V., Boullé, M., Dror, G., & Vogel, D. (2009, December). Analysis of the kdd cup 2009: Fast scoring on a large orange customer database. In *KDD-Cup 2009 Competition* (pp. 1-22).

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524.

Kotler, P. (1994). Marketing Management, Englewood Cliffs. *Jersey, USA*.

Li, H., Yang, D., Yang, L., & Lin, X. (2016, October). Supervised massive data analysis for telecommunication customer churn prediction. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)* (pp. 163-169). IEEE.

Lima, E., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: case studies in churn prediction. *Journal of the Operational Research*

Society, *60*(8), 1096-1106.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, *11*(3), 690-696.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547-12553.

Yu, R., An, X., Jin, B., Shi, J., Move, O. A., & Liu, Y. (2018). Particle classification optimization-based BP network for telecommunication customer churn prediction. *Neural Computing and Applications*, *29*(3), 707-720.

APPENDIX

```
/*---Student ID 30867835---*/
/*Before u run this code, it is better
to add a new lib name 'con'*/
/*Data collection*/
data con.ba1;
    infile'C:/Users/mktmi/OneDrive/桌面
/cons_sas/TelcoData_extract2.txt'
DLM=','
    DSD MISSOVER firstobs=2;
    input Customer_ID upsell_xsell
churn;
run;

PROC IMPORT OUT=con.ba2
DATAFILE="C:/Users/mktmi/OneDrive/桌面
/cons_sas/TelcoData_extract3.xlsx"
dbms=xlsx replace;
  getnames=yes;
run;

data con.ba;
merge con.ba1 con.ba2
con.TelcoData_extract1;
by Customer_ID;
run;

/*Descriptive Statistic for con.ba*/
proc means data=con.ba;
run;

/*pie chart*/
proc gchart data=con.ba;
pie network_mention  service_mention
    price_mention mfg_apple
    mfg_samsung   mfg_htc
    mfg_motorola
mfg_lg mfg_nokia  delinq_indicator
    times_delinq upsell_xsell   churn
    credit_class  sales_channel
region state  city   product_plan_desc
    handset_age_grp   handset
    lifestage  rp_pooled_ind
    call_center   issue_level1
issue_level2  call_category_1
    call_category_2   resolution
/type=sum percent=arrow slice=inside
ctext=black value=inside;
run;

/*bar chart*/
proc gchart data=con.ba;
hbar count_of_suspensions_6m
    avg_days_susp calls_total
    calls_in_pk   calls_in_offpk
    calls_out_offpk   calls_out_pk
voice_tot_bill_mou_curr
    tot_voice_chrgs_curr tot_drpd_pr1
    bill_data_usg_m03 bill_data_usg_m06
    bill_data_usg_m09
mb_data_usg_m01   mb_data_usg_m02
    mb_data_usg_m03
    mb_data_ndist_mo6m
    mb_data_usg_roamm01
    mb_data_usg_roamm02
    mb_data_usg_roamm03
data_usage_amt    tweedie_adjusted
    tot_mb_data_curr
    tot_mb_data_roam_curr
    bill_data_usg_tot tot_overage_chgs
data_prem_chrgs_curr nbr_data_cdrs
    avg_data_chrgs_3m
    avg_data_prem_chrgs_3m
    avg_overage_chrgs_3m nbr_contacts
calls_TS_acct open_tsupcomplnts
    num_tsupcomplnts
    unsolv_tsupcomplnt   wrk_orders
    days_openwrkorders
    resolved_complnts
calls_care_acct   calls_care_3mavg_acct
    calls_care_6mavg_acct
    res_calls_3mavg_acct
    res_calls_6mavg_acct
last_rep_sat_score lifetime_value
    avg_arpu_3m   acct_age
```

```
    billing_cycle nbr_contracts_ltd
    rfm_score  Est_HH_Income
zipcode_primary   region_lat
    region_long    state_lat  state_long
    city_lat   city_long  zip_lat
    zip_long   cs_med_home_value
cs_pct_home_owner cs_ttl_pop
    cs_hispanic    cs_caucsasian
    cs_afr_amer    cs_other
    cs_ttl_urban   cs_ttl_rural
cs_ttl_male    cs_ttl_female
    cs_ttl_hhlds   cs_ttl_mdage
forecast_region   mb_inclplan
    ever_days_over_plan
    ever_times_over_plan
    data_device_age   equip_age
/type=sum percent=arrow slice=inside
ctext=black value=inside;
run;

/*Remove duplicate values*/
/*0 duplicate values removed*/
proc sort data = con.ba out =con.cs
nodup;
   by Customer_ID;
run ;


/*Missing value query*/
data missing(drop=i);
set con.cs;
array a _numeric_;
do i=1 to dim(a);
if missing(a) then output;
end;

array b _character_;
do i=1 to dim(b);
if missing(b) then output;
end;

/*mou data set back up*/
data con.mou;
set con.cs;
if nmiss(of _numeric_) + cmiss(of
_character_) > 0 then delete;
run;

/*delete mou_ 6 column*/
data con.cs1;
    set con.cs;
    drop mou_total_pct_MOM
    mou_onnet_pct_MOM mou_roam_pct_MOM
    mou_onnet_6m_normal
    mou_roam_6m_normal call_category_2;
run;

/*Delete missing value*/
/*remain 46076 variables*/
data con.cs2;
set con.cs1;
if nmiss(of _numeric_) + cmiss(of
_character_) > 0 then delete;
if days_openwrkorders = '**' then
delete;
run;

/*Collinearity (Just have a
look),Spearman in the following*/
proc corr data = con.cs2;
var Customer_ID upsell_xsell churn
lifetime_value avg_arpu_3m acct_age
billing_cycle nbr_contracts_ltd
    rfm_score Est_HH_Income
zipcode_primary region_lat region_long
state_lat state_long city_lat
city_long
    zip_lat zip_long cs_med_home_value
cs_pct_home_owner cs_ttl_pop
cs_hispanic cs_caucasian  cs_afr_amer
    cs_other cs_ttl_urban  cs_ttl_rural
cs_ttl_male cs_ttl_female cs_ttl_hhlds
cs_ttl_mdage
    forecast_region mb_inclplan
ever_days_over_plan
ever_times_over_plan data_device_age
equip_age
    mfg_apple mfg_samsung mfg_htc
```

```
mfg_motorola mfg_lg mfg_nokia
delinq_indicator times_delinq
    count_of_suspensions_6m
avg_days_susp calls_total calls_in_pk
calls_in_offpk calls_out_offpk
    calls_out_pk
voice_tot_bill_mou_curr
tot_voice_chrgs_curr tot_drpd_pr1
bill_data_usg_m03 bill_data_usg_m06
    bill_data_usg_m09 mb_data_usg_m01
mb_data_usg_m02 mb_data_usg_m03
mb_data_ndist_mo6m mb_data_usg_roamm01
    mb_data_usg_roamm02
mb_data_usg_roamm03 data_usage_amt
tweedie_adjusted tot_mb_data_curr
tot_mb_data_roam_curr
    bill_data_usg_tot tot_overage_chgs
data_prem_chrgs_curr nbr_data_cdrs
avg_data_chrgs_3m
avg_data_prem_chrgs_3m
    avg_overage_chrgs_3m nbr_contacts
calls_TS_acct open_tsupcomplnts
num_tsupcomplnts unsolv_tsupcomplnt
    wrk_orders days_openwrkorders
resolved_complnts calls_care_acct
calls_care_3mavg_acct
    calls_care_6mavg_acct
res_calls_3mavg_acct
res_calls_6mavg_acct
last_rep_sat_score network_mention
service_mention price_mention;
run;


/*PCA*/
proc princomp data=con.cs2
out=con.prin prefix=z standard;
var Customer_ID upsell_xsell churn
lifetime_value avg_arpu_3m acct_age
billing_cycle nbr_contracts_ltd
    rfm_score Est_HH_Income
zipcode_primary region_lat region_long
state_lat state_long city_lat
city_long
    zip_lat zip_long cs_med_home_value
cs_pct_home_owner cs_ttl_pop
cs_hispanic cs_caucasian  cs_afr_amer
    cs_other cs_ttl_urban  cs_ttl_rural
cs_ttl_male cs_ttl_female cs_ttl_hhlds
cs_ttl_mdage
    forecast_region mb_inclplan
ever_days_over_plan
ever_times_over_plan data_device_age
equip_age
    mfg_apple mfg_samsung mfg_htc
mfg_motorola mfg_lg mfg_nokia
delinq_indicator times_delinq
    count_of_suspensions_6m
avg_days_susp calls_total calls_in_pk
calls_in_offpk calls_out_offpk
    calls_out_pk
voice_tot_bill_mou_curr
tot_voice_chrgs_curr tot_drpd_pr1
bill_data_usg_m03 bill_data_usg_m06
    bill_data_usg_m09 mb_data_usg_m01
mb_data_usg_m02 mb_data_usg_m03
mb_data_ndist_mo6m mb_data_usg_roamm01
    mb_data_usg_roamm02
mb_data_usg_roamm03 data_usage_amt
tweedie_adjusted tot_mb_data_curr
tot_mb_data_roam_curr
    bill_data_usg_tot tot_overage_chgs
data_prem_chrgs_curr nbr_data_cdrs
avg_data_chrgs_3m
avg_data_prem_chrgs_3m
    avg_overage_chrgs_3m nbr_contacts
calls_TS_acct open_tsupcomplnts
num_tsupcomplnts unsolv_tsupcomplnt
    wrk_orders days_openwrkorders
resolved_complnts calls_care_acct
calls_care_3mavg_acct
    calls_care_6mavg_acct
res_calls_3mavg_acct
res_calls_6mavg_acct
last_rep_sat_score network_mention
service_mention price_mention;
run;
```

```
/*Normality test*/
proc univariate data = con.prin normal
plot;
    var z1-z8;
run;

/*standardization*/
proc standard data=con.cs2 out=con.std
mean=0 std=1;
var Customer_ID upsell_xsell churn
lifetime_value avg_arpu_3m acct_age
billing_cycle nbr_contracts_ltd
    rfm_score Est_HH_Income
zipcode_primary region_lat region_long
state_lat state_long city_lat
city_long
    zip_lat zip_long cs_med_home_value
cs_pct_home_owner cs_ttl_pop
cs_hispanic cs_caucasian  cs_afr_amer
    cs_other cs_ttl_urban  cs_ttl_rural
cs_ttl_male cs_ttl_female cs_ttl_hhlds
cs_ttl_mdage
    forecast_region mb_inclplan
ever_days_over_plan
ever_times_over_plan data_device_age
equip_age
    mfg_apple mfg_samsung mfg_htc
mfg_motorola mfg_lg mfg_nokia
delinq_indicator times_delinq
    count_of_suspensions_6m
avg_days_susp calls_total calls_in_pk
calls_in_offpk calls_out_offpk
    calls_out_pk
voice_tot_bill_mou_curr
tot_voice_chrgs_curr tot_drpd_pr1
bill_data_usg_m03 bill_data_usg_m06
    bill_data_usg_m09 mb_data_usg_m01
mb_data_usg_m02 mb_data_usg_m03
mb_data_ndist_mo6m mb_data_usg_roamm01
    mb_data_usg_roamm02
mb_data_usg_roamm03 data_usage_amt
tweedie_adjusted tot_mb_data_curr
tot_mb_data_roam_curr
    bill_data_usg_tot tot_overage_chgs
data_prem_chrgs_curr nbr_data_cdrs
avg_data_chrgs_3m
avg_data_prem_chrgs_3m
    avg_overage_chrgs_3m nbr_contacts
calls_TS_acct open_tsupcomplnts
num_tsupcomplnts unsolv_tsupcomplnt
    wrk_orders days_openwrkorders
resolved_complnts calls_care_acct
calls_care_3mavg_acct
    calls_care_6mavg_acct
res_calls_3mavg_acct
res_calls_6mavg_acct
last_rep_sat_score network_mention
service_mention price_mention;
run;


/*PCA*/
proc princomp data=con.std
out=con.stdprin prefix=z standard;
var Customer_ID upsell_xsell churn
lifetime_value avg_arpu_3m acct_age
billing_cycle nbr_contracts_ltd
    rfm_score Est_HH_Income
zipcode_primary region_lat region_long
state_lat state_long city_lat
city_long
    zip_lat zip_long cs_med_home_value
cs_pct_home_owner cs_ttl_pop
cs_hispanic cs_caucasian  cs_afr_amer
    cs_other cs_ttl_urban  cs_ttl_rural
cs_ttl_male cs_ttl_female cs_ttl_hhlds
cs_ttl_mdage
    forecast_region mb_inclplan
ever_days_over_plan
ever_times_over_plan data_device_age
equip_age
    mfg_apple mfg_samsung mfg_htc
mfg_motorola mfg_lg mfg_nokia
delinq_indicator times_delinq
    count_of_suspensions_6m
avg_days_susp calls_total calls_in_pk
calls_in_offpk calls_out_offpk
    calls_out_pk
```

```
voice_tot_bill_mou_curr
tot_voice_chrgs_curr tot_drpd_pr1
bill_data_usg_m03 bill_data_usg_m06
    bill_data_usg_m09 mb_data_usg_m01
mb_data_usg_m02 mb_data_usg_m03
mb_data_ndist_mo6m mb_data_usg_roamm01
    mb_data_usg_roamm02
mb_data_usg_roamm03 data_usage_amt
tweedie_adjusted tot_mb_data_curr
tot_mb_data_roam_curr
    bill_data_usg_tot tot_overage_chgs
data_prem_chrgs_curr nbr_data_cdrs
avg_data_chrgs_3m
avg_data_prem_chrgs_3m
    avg_overage_chrgs_3m nbr_contacts
calls_TS_acct open_tsupcomplnts
num_tsupcomplnts unsolv_tsupcomplnt
    wrk_orders days_openwrkorders
resolved_complnts calls_care_acct
calls_care_3mavg_acct
    calls_care_6mavg_acct
res_calls_3mavg_acct
res_calls_6mavg_acct
last_rep_sat_score network_mention
service_mention price_mention;
run;

/*Normality test*/
/*The result is still not following
the normal distribution*/
proc univariate data = con.stdprin
normal plot;
    var z1-z8;
run;

/*Convert character variables to
numeric variables*/
data con.cs3(drop=i);
set con.cs2;

    array sample{7} credit_class
sales_channel region handset_age_grp
handset    lifestage rp_pooled_ind;
    do i=1 to 7;
            if sample{i}= 'risky'
then sample{i}=1;
       else if sample{i}= 'other'
then sample{i}=2;
       else if sample{i}= 'near prime'
then sample{i}=3;
       else if sample{i}= 'prime'
then sample{i}=4;
       else if sample{i}= 'smax prime'
then sample{i}=5;

            if sample{i}= 'Direct'
         then sample{i}=1;
       else if sample{i}= 'Private
Label GM'        then sample{i}=2;
       else if sample{i}= 'Indirect'
            then sample{i}=3;
       else if sample{i}= 'Branded 3rd
Party Retail' then sample{i}=4;
       else if sample{i}= 'Retail'
            then sample{i}=5;
       else if sample{i}= 'National
Sales'      then sample{i}=6;

            if sample{i}= 'Great Lakes'
then sample{i}=1;
       else if sample{i}= 'Greater
Texas' then sample{i}=2;
       else if sample{i}= 'Mid
Atlantic'  then sample{i}=3;
       else if sample{i}= 'Midwest'
then sample{i}=4;
       else if sample{i}= 'Mtn West'
then sample{i}=5;
       else if sample{i}= 'New
England'   then sample{i}=6;
       else if sample{i}= 'Pacific'
then sample{i}=7;
       else if sample{i}= 'South'
then sample{i}=8;
       else if sample{i}= 'Southwest'
then sample{i}=9;
```

```
           if sample{i}= '< 24 Months'
then sample{i}=1;
       else if sample{i}= '24-48
Month' then sample{i}=2;
       else if sample{i}= '> 48
Months' then sample{i}=3;

           if sample{i}= 'Apple'
then sample{i}=1;
       else if sample{i}= 'HTC'
then sample{i}=2;
       else if sample{i}= 'LG'
then sample{i}=3;
       else if sample{i}= 'Motorola'
then sample{i}=4;
       else if sample{i}= 'Nokia'
then sample{i}=5;
       else if sample{i}= 'Samsung'
then sample{i}=6;
       else if sample{i}= 'Unknown'
then sample{i}=7;

           if sample{i}= 'EARLY
TENURED' then sample{i}=1;
       else if sample{i}= 'EXPIRY'
then sample{i}=2;
       else if sample{i}= 'OFF-
CONTRACT'  then sample{i}=3;
       else if sample{i}= 'ON-
CONTRACT'   then sample{i}=4;
       else if sample{i}= 'PRE-EXPIRY'
then sample{i}=5;

           if sample{i}= 'N'      then
sample{i}=0;
       else if sample{i}= 'Y'
then sample{i}=1;

   end;
run;

/*Numeric Outliers*/
PROC UNIVARIATE DATA=con.cs3 plot;
VAR _NUMERIC_;
RUN;

/*Outliers*/
data con.cs4;
set con.cs3;
if mb_data_usg_m03 = -509 then delete;
if mb_data_usg_roamm02 = 18727 then
delete;
if mb_data_usg_m03 = 40784 then
delete;
run;

/*Deduplication*/
data con.cs5;
   set con.cs4;
   drop mfg_samsung mfg_nokia
mfg_motorola mfg_lg mfg_htc mfg_apple;
run;

/*Pearson coefficient matrix*/
proc corr spearman nosimple
data=con.cs5;
   var _NUMERIC_;
run;

/*Remove multicollinearity, judge by
Pearson coefficient > 0.8*/
data con.cs6;
   set con.cs5;
   drop region_long state_long
city_long zip_long forecast_region
cs_ttl_hhlds
       cs_ttl_rural cs_ttl_female
res_calls_3mavg_acct data_usage_amt
mb_data_usg_roamm01
       mb_data_usg_roamm02
mb_data_usg_roamm03 calls_in_pk
calls_in_offpk calls_out_pk
       voice_tot_bill_mou_curr
res_calls_6mavg_acct;
run;

data con.cs7;
   set con.cs6;
```

```
       credit_class_n =
credit_class*1;
       sales_channel_n =
sales_channel*1;
       region_n = region *1;
       handset_age_grp_n
=handset_age_grp*1;
       handset_n  = handset*1;
       lifestage_n = lifestage*1;
       rp_pooled_ind_n =
rp_pooled_ind*1;
       credit_class_n =
credit_class*1;
    drop credit_class sales_channel
region handset_age_grp handset
lifestage rp_pooled_ind;
run;


/*Individual Part*/
proc sort data = con.cs7;
by Churn;
run;

/*Sampling for solving data
imbalance*/
PROC SURVEYSELECT DATA = con.cs7 out =
con.samp1 method = srs sampsize = 2000
seed = 123;
     STRATA Churn;
RUN;

/*Out of order*/
data con.samp2;
    set con.samp1;
    rand = uniform(12);
run;
proc sort data = con.samp2;
by rand;
run;
data con.samp3;
    set con.samp2;
    drop rand Customer_ID upsell_xsell
call_center issue_level1 issue_level2
call_category resolution
        state city product_plan_desc
call_category_1 SelectionProb
SamplingWeight product_plan_desc;
run;

data con.samp4;
    set con.samp3;
    drop count_of_suspensions_6m
calls_care_acct price_mention
        tot_drpd_pr1 nbr_contacts
last_rep_sat_score;
run;

/*Inspection data type*/
proc contents data=con.samp4 out=a;
run;


/*Change the variable name to make it
easy to code*/
data con.samp5;
  set con.samp4;
  array x{69} churn  lifetime_value
avg_arpu_3m  acct_age  billing_cycle
nbr_contracts_ltd  rfm_score
Est_HH_Income  zipcode_primary
region_lat  state_lat  city_lat
zip_lat
            cs_med_home_value
cs_pct_home_owner  cs_ttl_pop
cs_hispanic  cs_caucasian  cs_afr_amer
cs_other  cs_ttl_urban  cs_ttl_male
cs_ttl_mdage  mb_inclplan
            ever_days_over_plan
ever_times_over_plan  data_device_age
equip_age  delinq_indicator
times_delinq    avg_days_susp
calls_total
            calls_out_offpk
tot_voice_chrgs_curr bill_data_usg_m03
bill_data_usg_m06  bill_data_usg_m09
mb_data_usg_m01  mb_data_usg_m02
```

```
            mb_data_usg_m03
mb_data_ndist_mo6m  tweedie_adjusted
tot_mb_data_curr  tot_mb_data_roam_curr
bill_data_usg_tot  tot_overage_chgs
data_prem_chrgs_curr
            nbr_data_cdrs
avg_data_chrgs_3m
avg_data_prem_chrgs_3m
avg_overage_chrgs_3m  calls_TS_acct
open_tsupcomplnts  num_tsupcomplnts
            unsolv_tsupcomplnt
wrk_orders  days_openwrkorders
resolved_complnts
calls_care_3mavg_acct
calls_care_6mavg_acct
            network_mention
service_mention  credit_class_n
sales_channel_n  region_n
handset_age_grp_n  handset_n
lifestage_n  rp_pooled_ind_n
            ;
  array v{69} v1-v69;
  do i=1 to 69;
   v{i}=x{i};
     end;
  keep v:;
run;

data con.samp6;
    set con.samp5;
    Y=v1;
    drop v1;
run;


/*Using 10-fold cross-validation
method to calculate the prediction
accuracy on the test set*/
%let k=10;
%let rate=%sysevalf((&k-1)/&k);

/*Generate 10 examples of cross-
validation, save in cv*/
proc surveyselect data=con.samp6
                  out=cv
                  seed=158
                  samprate=&rate
                  outall
                  reps=10;
run;


data cv;
  set cv;
   if selected then new_y=Y;
  run;

/*Logistic regression main program-10%
off cross-validation*/
ods output parameterestimates=paramest
          association=assoc;
proc logistic data=cv des;
       model new_y = v2-v69 /
SELECTION=STEPWISE SLE=0.1 SLS=0.1;
    by replicate;
    output out=out1(where=(new_y=.))
          p=y_hat;
run;
ods output close;

data out1;
    set out1;
    if y_hat>0.5 then pred=_LEVEL_ ;
    else pred=0;
run;

/*Summarize the results of cross-
validation*/
data out2;
    set out1;
    if Y=pred then d=1;
    else d=0;
run;

proc summary data=out2;
    var d;
    by replicate;
    output out=out3 sum(d)=d1;
```

```
run;

data out3;
    set out3;
    acc=d1/_freq_;
    keep replicate acc;
run;

/*Include cross-validated C statistics
in the results*/
data assoc;
    set assoc;
    where label2="c";
    keep replicate cvalue2;
run;

/*Combine the statistical results of
cross-validation*/
data cvresult;
merge assoc(in=ina) out3(in=inb);
keep replicate cvalue2 acc;
run;

proc print data=cvresult;
title'Cross-validation group number, c
statistics, prediction accuracy';
run;

title 'Cross-validation optimal model
selection: group number, prediction
accuracy';
ods output SQL_Results=cvparam;
proc sql ;
    select replicate,acc from cvresult
having acc=max(acc);
quit;
ods output close;



/***************** Model with cross-
validated optimal result set
************************************/
proc sql ;
    create table train as
    select * from cv where replicate in
(select replicate from cvparam)
    having selected=1;
    create table test as
    select * from cv where replicate in
(select replicate from cvparam)
    having selected=0;
run;

TITLE '--------Logistic Regression----
--------';

/* Logistic regression main program-
build a logistic model from the
training set*/
proc logistic data=train  DES
                  covout
outest=Nout_step
                   outmodel=model
                  simple;
       MODEL Y=v2-v69
                    / SELECTION=STEPWISE
                      SLE=0.1 SLS=0.1
                      details
                      lackfit
                      RSQ
                       STB
                      CL
                      itprint
                      corrb
                      covb
                      ctable
                      influence
                       IPLOTS ;
 score data=train outroc=train_roc;
 score data=test
      out=test_pred
      outroc=test_roc;
OUTPUT out=train_pred
          P=PHAT  lower=LCL upper=UCL
          RESCHI=RESCHI  RESDEV=RESDEV
          DIFCHISQ=DIFCHISQ
```

```
DIFDEV=DIFDEV

          / ALPHA=0.1;
run;
quit;



data train_pred;
    set train_pred;
    if PHAT>0.5 then pred=_LEVEL_ ;
    else pred=0;
run;

/* Output confusion matrix-training
set*/
ods output CrossTabFreqs=ct_train;
ods trace on;
proc freq data=train_pred;
    tables Y*pred;
run;
ods trace off;
ods output close;

proc sql;
    create table acc1 as
    select sum(percent) from ct_train
where (Y=pred and Y ^=.);
proc print data=acc1;
title 'Prediction accuracy on the
training set';
run;



/* Output indicators such as confusion
matrix and accuracy-test set*/
ods output CrossTabFreqs=ct_test;
proc freq data=test_pred;
    tables F_Y*I_Y ;
run;
ods output close;

proc sql;
    create table acc2 as
    select sum(percent) from ct_test
where (F_Y=I_Y and F_Y ^='');
proc print data=acc2;
title 'Prediction accuracy on test
set';
run;

/*Clustering*/
proc gchart data=con.samp3;
hbar avg_overage_chrgs_3m
/type=sum ;
run;

data con.km;
    set con.samp3;
    keep avg_overage_chrgs_3m;
run;

proc fastcluster data=con.km maxc=3
maxiter=10 list output=con.c11;
var avg_overage_chrgs_3m;
run;
```