Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression

Peng Li [1, 2], Siben Li [2], Tingting Bi [2], Yang Liu [2]

[1] *School of Software, Harbin University of Science and Technology, 150080 Harbin, China*
[2] *School of Computer and Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China*
pli@hrbust.edu.cn.

Keywords: imbalance data; customer churn; stratified sampling; logistic regression model.

Abstract

This paper provides a novel and efficient method for predicting potential customer churn from imbalanced data set of Orange Telecom and UCI. Customer churn is always a rare event, but it is necessary to be paid attention. The main intended contribution of this paper is to apply binary logistic regression model (LRM), which seldom be used in the problem of imbalanced data prediction. The parameters estimated for the data gathered with serious problem of imbalance, therefore we take stratified sampling method, and improve traditional logistic regression model parameters estimated methods. The experimental results show that our prediction method performs satisfactorily, and it can be effective to forecast the telecom customer churn.

1 Introduction

Customer churn prediction has recently received more and more attention by reason of advances in the areas of data mining, statistical learning and business management etc. Customer relationship management (CRM) is a mean of corporation to improve their competitiveness. Customer churn prediction plays an important role in CRM. Customers are very unstable group. Enterprises, of course, are happy to retain them. Customer retention is the most significant issues for telecom companies. Customer churn prevention is on the agenda.

People often keep using one telephone code during several years due to the objective requirement of communication. Sometimes it can last even dozen of years. But in CRM, the cost of a new customer is five times that of retaining an old customer. For reducing customer churn by 5%, profits can be increased by 25%~85% [1]. A satisfied customer will bring 8 potential deals to enterprises. While an unsatisfied customer likely to affect the purchase intention of 25 persons. Most companies will churn half of their customers in 5 years, if they ignore old customers [2]. It is thus clear that how important customers churn is in telecom companies. Therefore, many researchers around the world focus on this topic and proposed some methods and strategies to deal with this problem, such as Idris employed mRMR feature selection and RotBoost ensemble learning to predict churn in telecom [3]; Lu proposed a customer churn prediction model in telecom industry using boosting [4]; Li applied cluster analysis and decision tree algorithm to forecast customer churn of China Telecom[5]; Verbraken predict customer churn by mean of Bayesian network classifier [6] and so on. Unfortunately, the data imbalance problem of customer churn is rarely mentioned.
Because of the telecom customer chum is often a rare event, but of great interest and great value [7]. In data mining, this is a typical imbalanced dataset classification problem. For example, in financial fraud detection, the vast majority of users are all legitimate users, but people are very hope that through data to predict the potential illegal user [8]; In bankruptcy risk prediction, the bankrupt company is minority of all, but the managers also care about the present management state whether bankruptcy is possible [9]; In medical diagnosis, real data must be the most healthy people, and people care about is whether through the existing data to predict the occurrence of diseases [10]; In general, customer churn is often a rare event. There is an ocean of data generated, when telecom companies are running. But special part accounts for little of the customer relationship data, but is of great interest and great value to customer churn analysis. To dig this part out is a top priority in customer chum prediction. In data mining, this kind of data form is called imbalanced data.

In this article, we proposed an innovative method emphatically, namely, cluster stratified sampling logistic regression model (CSS-LRM), solve the problem of imbalanced data in customer churn prediction.

2 Stratified sampling based on K-means clustering

Clustering is one of the most common techniques for data mining, for the discovery of unknown data types in the database, used to find the unknown data types in the database, formed by the clustering process of each group is called a class. Before the clustering, the number and type of

data is unknown. The data classification is based on the "Like attracts like", according to the individual or the similarity between data objects, the research object is divided into several categories. Cluster, a group of objects according to similarity into several categories, is to make them belong to the same class of objects as similarity as possible features, and belongs to a relative independence as possible between different classes of objects. From the above described about clustering，we found much in common and intersection in the guiding ideology of clustering theory and hierarchy. Therefore, the clustering method provided a good theory basis and feasible method for dividing layers of stratified sampling.

At present, there are many existing clustering methods, such as: segmental clustering, hierarchical clustering, density clustering etc. Among them, the K-means clustering algorithm is a simple, effective and easy to control and improvement has become the most widely used clustering algorithm. It has been successfully applied in many research fields, such as: image processing, network optimization, natural language processing and so on.
This chapter will divide the K-means clustering algorithm is applied in layers in stratified sampling, the K-means clustering algorithm in addition to its simple, effective features, the most important is, the number of the clustering algorithm to cluster categories can be set in advance. From the hierarchy, the application of this algorithm is also can be defined prior to divide layer, it can effectively control the sampling process. Suppose that there are *N* objects need to be divided into class *K*, then in k-means algorithm, first select the *K* object representing the *K* classes randomly, each object as a center, according to the nearest to the center principle will be assigned to each class of other objects. After the allocation of the completion of the first object, to each attribute means all objects in each class as the class of the new center, the redistribution of objects, the process is repeated until no change so far, to get the final class *K*. In K-means algorithm, cluster number *K*, is a parameter must be specified in advance. The clustering process can be described by the following steps:

(1) Randomly selected K objects, each object as a kind of "center", representing the will be divided into k classes;
(2) According to the distance from the "center" of recent principle, the other objects assigned to each corresponding class;
(3) For each class. Calculation of the average properties of all the value of the object, as the new "center";
(4) According to the distance from the "center" of recent principle, to all objects to allocate the corresponding class;
(5) If (4) divided by new class and original class division the same, then stop the calculation. Otherwise, go to (3).

3 Customer churn prediction based on logistic regression model with parameter compensation

Logistic regression model (LRM) is a regular and effective method of statistical analysis for two-category regression analysis. It has extensive application in such fields as economics [11], sociology [12], and medicine[13] etc, but it is less in the field of information processing. Logistic regression is a nonlinear model ， therefore the parameters of the model are estimated by maximum likelihood generally. It is proved that maximum-likelihood estimation of logistic regression has the characteristics of consistency, asymptotic validity and asymptotic normality. Maximum-likelihood estimation methods have a number of attractive attributes. First, they nearly always have good convergence properties as the number of training samples increases. Furthermore, maximum-likelihood estimation often can be simpler than alternative methods, such as Bayesian techniques or other methods.

Customer churn prediction is typical two-category case, because one candidate customer only has two kinds of situations, that it is a churn or not. Therefore, this kind of problem is suitable for the method of logistic regression for analyzing. But in the actual conditions, the positive instance (churn) far less than negative instance (no churn), it brings about serious data imbalance. In this case, if you directly adopt maximum-likelihood estimation, it will result in the model parameter and probability estimate deviation. This paper brings forward a method of parameter estimation, which can diminish the deviation of estimation.

3.1 Binary Logistic Regression: Model and Parameter Estimated

In logistic regression, a single outcome variable $Y_i \ \ (i = 1, 2, 3, ..., n)$ follows a Bernoulli probability function that takes on the value 1 with probability P_i and 0 with probability $1 - P_i$. $P_i / 1 - P_i$ is referred to as the $odds$ of an event occurring. Then P_i varies over the observations as an inverse logistic function of a vector $X_i \ \ (i = 1, 2, 3, ..., n)$ which includes a constant and K explanatory variables:

$$Y_i \sim Bernoulli(Y_i / P_i) \tag{1}$$

$$\ln \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \ln(odds) = \alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik} \tag{2}$$

The above is referred to as the log odds and also the logit. By taking the antilog of both sides, the model can also be expressed in odds rather than log odds, i.e.

$$odds = \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \exp\left(\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik} \right) \tag{3}$$

$$= e^{\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik}} = e^{\alpha_0} * \prod_{k=1}^{K} e^{\beta_k X_k} = e^{\alpha_0} * \prod_{k=1}^{K} (e^{\beta_k})^{X_k} \tag{4}$$

As Aldrich and Nelson note, there are several alternatives to the LRM that might be just as plausible or more plausible in a particular case. However, the LRM is comparatively easy from a computational standpoint; there are many tools available

which can estimate logistic regression models and the LRM tends to work fairly well in practice.

Note that, if we know either the odds or the log odds, it is easy to figure out the corresponding probability:

$$P_{x_i} = \frac{odds}{1+odds} = \frac{\exp\left(\alpha_0 + \beta^{'} X\right)}{1+\exp\left(\alpha_0 + \beta^{'} X\right)} \tag{5}$$

The unknown parameter α_0 is a scalar constant term and $\beta^{'}$ is a $K \times 1$ vector with elements corresponding to the explanatory variables. The parameters of the model are estimated by maximum likelihood. That is, the coefficients that make our observed results most "likely" are selected. The likelihood function formed by assuming independence over the observations:

$$L\left(\alpha_0, \beta^{'}\right) = \prod_{i=1}^{n} P_{x_i}^{Y_i} \left(1 - P_{x_i}\right)^{1-Y_i} \tag{6}$$

To random sample (x_i, y_i) ,$i = 1, 2, ..., n$ By taking logs and using formula (2) the log-likelihood simplifies to

$$\ln\left(L\left(\alpha_0, \beta^{'}\right)\right) = \sum_{i=1}^{n}\left[y_i(\alpha_0 + \beta^{'} x_i) - \ln\left(1+\exp\left(\alpha_0 + \beta^{'} x_i\right)\right)\right] \tag{7}$$

The estimator of unknown parameter α_0 and $\beta^{'}$ can be gained from following equations by means of maximum-likelihood estimation.

$$\begin{cases} \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \alpha_0} = \displaystyle\sum_{i=1}^{n}\left| y_i - \frac{\exp(\alpha_0 + \beta^{'} x)}{1+\exp(\alpha_0 + \beta^{'} x)}\right| = 0 \\ \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \beta_j} = \displaystyle\sum_{i=1}^{n}\left| y_i - \frac{\exp(\alpha_0 + \beta^{'} x)}{1+\exp(\alpha_0 + \beta^{'} x)}\right| x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{cases} \tag{8}$$

3.2 Stratified Sampling Logistic Regression in Imbalanced Data

In actual application, it often have lager gap between the positive instance and the negative instance, and the positive instance far less than negative instance, so such data have serious data sparse problem. If we adopt general logistic regression to estimate parameters in such data, usually the results are not good or even the wrong. Therefore, we utilized the method of stratified sampling to take full advantage of the resource of positive instances. The concrete process is: random extract some examples from positive instances and negative instances and merge the training samples to parameter estimation.

Under the condition of stratified sampling, sample distribution and population distribution doesn't have identity. In other word, the conditional probability of a sample observed value can't be expressed by formula (6) and formula (8) can't be found naturally.

Assuming that positive instances and negative instances have $P_0 N$ and $(1-P_0)N$ respectively among the population, the positive instances of independent variable x divided by total positive instances is γ_x , then the positive instances of independent variable x is $P_0 N \gamma_x$. We assume that the negative instances of independent variable x is κ_x , namely,

$$P_x = P_0 N \gamma_x / (P_0 N \gamma_x + \kappa_x) \tag{9}$$

Then, $\kappa_x = (1-P_x) P_0 N \gamma_x / P_x$ and the negative instances of independent variable x divided by total negative instances is λ_x.

$$\lambda_x = (1-P_x)\gamma_x P_0 / (1-P_0) P_x \tag{10}$$

Adopting the method of stratified sampling, we randomly extract γ_1 positive instances and γ_2 negative instances as sample. The probability of the observed value $y = 1$, $y = 0$ is:

$$P_x(1) = \frac{r_1 \gamma_x}{r_1 \gamma_x + r_2 \lambda_x} = \frac{r_1 (1-P_0) P_x}{r_1 (1-P_0) P_x + r_2 (1-P_x) P_0} \tag{11}$$

$$P_x(0) = \frac{r_2 \lambda_x}{r_1 \gamma_x + r_2 \lambda_x} = \frac{r_2 (1-P_x) P_0}{r_1 P_x (1-P_0) + r_2 (1-P_x) P_0} \tag{12}$$

Assuming $\omega_0 = P_0 N / (1-P_0) N$, $\omega_1 = r_1 / r_2$, namely, ω_0 is the ratio of the positive instances and the negative instances in population; ω_1 is the ratio of the positive instances and the negative instances in sample. As to stratified sample (x_i, y_i) , $i = 1, 2, ...n$, the logarithmic likelihood function is:

$$\begin{aligned} \ln[L(\alpha_0, \beta)] &= \sum_{i=1}^{n} \left\{ \begin{array}{c} y_1 (\ln \omega_1 + \ln P_{x_i}) + \\ (1-y_i)[\ln \omega_0 + \ln(1-P_{x_i})] - \\ \ln[\omega_1 P_{x_i} + (1-P_{x_i})\omega_0] \end{array} \right\} \\ &= \sum_{i=1}^{n} | \, y_i \ln \frac{\omega_1}{\omega_0} | + \sum_{i=1}^{n} | \, y_i \ln \frac{P_{x_i}}{1-P_{x_i}} | - \sum_{i-1}^{n} | \frac{\omega_1}{\omega_0} \frac{P_{x_i}}{1+P_{x_i}} + 1 | \end{aligned} \tag{13}$$

Utilizing formula (2), the log-likelihood simplifies to

$$\ln[L(\alpha_0, \beta)] = \Omega + \sum_{i=1}^{n} \left\{ \begin{array}{c} y_i (\alpha_0 + \beta^{'} x_i) - \\ \ln[1+\exp(\alpha_0 + \omega + \beta^{'} x_i)] \end{array} \right\} \tag{14}$$

Here, $\Omega = \omega \sum_{i=1}^{n} y_i$ and $\omega = \ln \omega_1 / \omega_0$ are nothing to estimated parameters. If we assume that $\alpha_1 = \alpha_0 + \omega$, then the estimator of unknown parameter α_1 and $\beta^{'}$ can be gained from following equations by means of maximum-likelihood estimation.

$$\left\{ \begin{array}{l} \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \alpha_0} = \displaystyle\sum_{i=1}^{n} | \, y_i - \frac{\exp(\alpha_1 + \beta^{'} x)}{1+ex\,\mathrm{p}(\alpha_1 + \beta^{'} x)} | = 0 \\ \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \beta_j} = \displaystyle\sum_{i=1}^{n} | \, y_i - \frac{\exp(\alpha_1 + \beta^{'} x)}{1+ex\,\mathrm{p}(\alpha_1 + \beta^{'} x)} | \, x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{array} \right\} \tag{15}$$

Formula (15) is the parameter estimation formula of stratified sampling logistic regression model. Under the condition of random sampling, sample distribution is identical to population distribution, $\omega_1=\omega_0$, then $\omega=0$, $\alpha_1=\alpha_0$, formula (15) is equal to formula (8). Therefore, formula (15) can be considered as an expansion of formula (8) under the condition of stratified sampling.

4 Experiment and analysis

In order to solve the problem of telecom customer churn prediction, we select two data sets (Churn and Orange) in the experiment. Chum set is a dataset representing telecom customer churn in UCI Machine Learning Repository. The dataset is satisfactory that the dataset has no missing value and don't need too much data preprocessing. Orange dataset is from Orange Telecom. In this dataset, there are excessive missing values, dimensions and complex information. Mountains of data preprocessing work should be done before classification.

The two above datasets can be used to test and verify cluster-based boundary sampling method that improving the availability of imbalanced data classification in scientific research and practical application. The information of the two data sets is shown in Table 1.

Data Set	Churn	Orange
Missing Value	No	Yes
Characteristics	Multivariate	Multivariate
Number of Attributes	20	230
Number of Instances	5000	50000
Imbalanced Ratio	6:1	13:1

Table 1. Information of churn and orange

4.1 Evaluation Metrics of Customer Churn

Due to the distribution imbalance of dataset type, commonly used evaluation metrics is less strong in the evaluation of imbalanced data classification performance. In recent years, many researches show that using ROC curve and AUC has obvious advantages to evaluate the performance of the imbalance data set classification, because they cannot be affected by the distribution imbalance of data type. That means ROC curve and AUC will not change when the number of positive and negative of the test data is changed. They can be scientific and intuitive to evaluate the performance of the classification [14].

ROC curve is a two-dimensional curve, horizontal indicates FPR (False Positive Rate), the vertical indicates TPR (True Positive Rate).The more test data, the smoother ROC curve is. In Figure. 1, if curve X is located above the curve Y. We consider curve X is better than Y. That is to say the expectation cost of classifier X is always lower than Y's for all possible misclassification cost and class distribution.

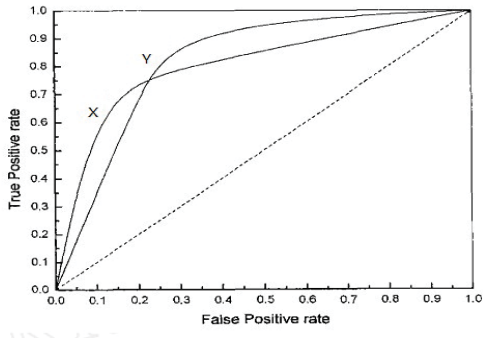

Figure.1. Two intersecting curve of ROC

Although ROC curve can show the performance of classification intuitively, we also want to use numerical description to evaluate classification result. As shown in figure 5, if two ROC curves intersect, we can just intuitively found X is better than Y when FPR is less than 0.23.On the contrary, Y is better than X when FPR is greater than 0.23.If only choosing ROC, it is hard to say which is better between X and Y. We even cannot know the gap between them. But we can calculate the area under ROC curve (AUC values).So this problem is solved. Classification performance is showed intuitively and clearly.

When carrying on the classification task. The minority class with fewer samples is defined as positive, and the majority is defined as negative. The results are divided into four kinds of circumstances (TP , FP , FN , TN).Positive classified correctly is TP .Positive classified wrongly is FP .Negatives classified correctly is TN .And negatives classified wrongly is FN .So the number of virtual positive $P = TP + FN$,while the number of negative $N = TN + FP$.

$$AUC = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{P \cdot N} \int_0^N TPdFP \qquad (16)$$

4.2 The Experimental Results and Analysis

In this paper, we use two strategies to conduct an experiment, which is to verify that the proposed strategy plays an effective role on classification, on the above two data sets.

Method 1: Using SVM of common RBF kernel function model to classify;
Method 2: Using CSS-LRM with parameter compensation to predict.
AUC values of prediction results and ROC figure are shown in the following:

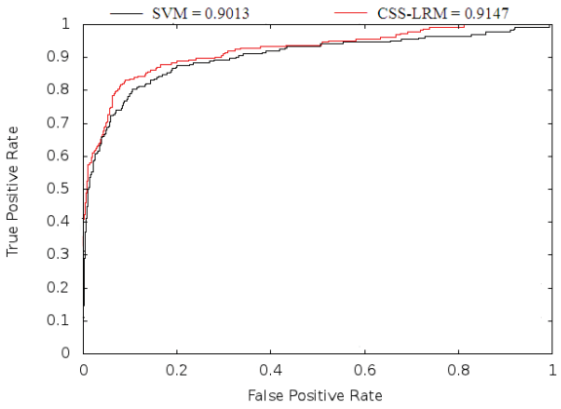

Figure.2. ROC curves of Churn

From Figure.2 and Figure.3, we can see ROC curve can intuitively describe the prediction performance. In order to further the quantitative comparison, we calculate AUC values of two methods according to the ROC curve, as shown in Table 2.

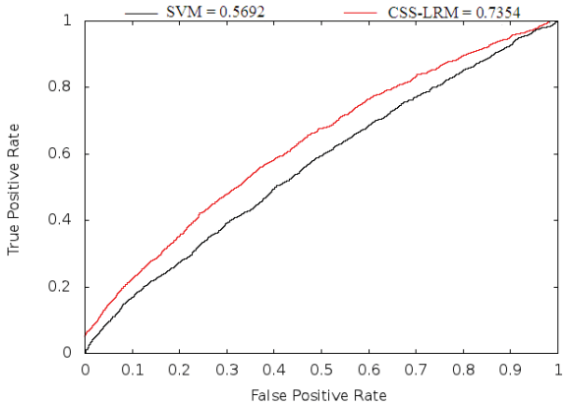

Figure.3. ROC curves of Orange

Data Set	SVM	CSS-LRM
Churn	0.9013	0.9147
Orange	0.5692	0.7354

Table 2. The AUC experimental results

SVM shows the good classification performance due to the lower Churn data set attribute dimension and imbalance ratio. However, the performance is declined slightly under the conditions of high dimension and imbalance ratio. The data of Orange dataset are real. It is relatively complex and affected by above uncontrollable factors. Our method raises the prediction results from0.5692 to 0.7354.

5 Conclusion

This paper proposed a customer churn prediction method based on cluster stratified sampling logistic regression model with parameters estimated methods that is suitable for imbalance data set. Using UCI and Orange to experiment on representative public data sets, with ROC curves and AUC value as the evaluation index of experiments, comparing the experimental results show that the presented method for telecom customer churn prediction has the stable promotion effect. In the future research work, according to the prediction of imbalanced data sets is still exist many problems. Imbalance data not only positive and negative cases of imbalance, mass data, incomplete data and data sparse will be further explored.

Acknowledgements

This paper is partially supported by National Natural Science F oundation of China (61103149), Postdoctoral Science Foundati on (2011M500682, LBHZ11106), Technological Innovation F oundation for Youth Scholars of Harbin (2012RFQXG093) an d Natural Science Foundation of Province (QC2013C060).

References

[1] H.F.Qin, "The Application of Data Mining in Telecommunication Churn Customer," *Research Journal of Applied Sciences*, vol.38,pp.1054-1057, 2012.

[2] A.Idris, A.Khan, Y.S.Lee ,"Genetic Programming and Adaboosting based churn prediction for Telecom," *IEEE Transactions on Systems, Man, and Cybernetics*,vol.20, pp. 1328-1332, 2012.

[3] A.Idris, A.Khan, Y.S.Lee, "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification,"*Applied Intelligence*,vol.39,pp.659-672,2013.

[4] N.Lu,H.Lin,J.Lu,G.Q.Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*,vol.10,pp.1659-1665,2014.

[5] G.Q.Li,X.Q.Deng, "Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm," *Communications in Computer and Information Science*,vol.315,pp.319-327,2012.

[6] Verbraken.T,Verbeke.W,Baesens.B, "Profit optimizing customer churn prediction with Bayesian network classifiers," *Intelligent Data Analysis*,vol.18,pp.3-24,2014.

[7] Z.Zhang,H.Lin,K.Liu,et al."A Hybrid Fuzzy-Based Personalized Recommender System for Telecom Products/Services," *Information Sciences*, vol.235,pp.117-129,2013.

[8] W.Wei,J.Li , L.Cao, et al. "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol.18,pp.1-27,2012.

[9] L.Zhou , "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Systems*, vol.41,pp.16-25,2013.

[10] J.Nahar, T.Imam, K.S.Tickle, et al. "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications*,vol.40,pp.96-104, 2012.

[11] G.L.Nie,R.Wei,L.L.Zhang,et al, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*,vol.38,pp.15273-15285,2011.

[12] G.King,M.Tomz,L.C.Zeng, "Relogit:Rare Events Logistic Regression," *Journal of Statistical Software*,vol.8,pp.84-113, 2003

[13] Vairavan.S,Eshelman.L,Haider.S,Flower.A,Seiver.A, "Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model," *Computing in Cardiology*,vol.39,pp-393-396,2012.

[14] J.Burez,D.Van den Poel, "Handling class imbalance in customer churn prediction ," *Expert Systems with Applications*,vol.36,pp.4626-4636,2009.