İstanbul Üniversitesi İşletme Fakültesi Dergisi
Istanbul University Journal of the School of Business Administration
Cilt/Vol:39, Sayı/No:1, 2010, 35-49
ISSN: 1303-1732 – www.ifdergisi.org © 2010

Customer churn analysis in telecommunication sector

Umman Tuğba Şimşek Gürsoy[1]
Department of Quantitative Methods,
School of Business Administration
Istanbul University, Istanbul, Turkey

Abstract

Data mining is used to analyze mass databases for having meaningful output. One of the most common applications of the data mining, which is called as Churn Analysis is used to predict behavior of customers who are most likely to change provided service, and to create special marketing tools for them. The aim of this paper is to determine customers who want to churn, and to create specific campaigns to them by using a customer data of a major telecommunication firm in Turkey. To determine the reasons of the customer churn, logistic regression and decision trees analysis, which is one of the classification techniques, are applied.

Keywords: *Churn analysis, data mining, logistic regression, decision trees, telecommunication.*

Telekomünikasyon sektöründe müşteri ayrılma analizi

Özet

Veri madenciliği, büyük veri kümeleri içindeki anlamlı bilgiyi ortaya çıkarma sürecidir. Veri madenciliğinin yaygın olarak kullanıldığı uygulama alanlarından biri, ayrılma eğilimi gösteren müşterilerin tahmin edilmesidir. Churn adı verilen bu analiz, şirketlerin kaybetme potansiyeli olan müşterilerine özel pazarlama kampanyalarını geliştirmelerini sağlamaya yöneliktir. Bu çalışma, Türkiye'de telekomünikasyon sektöründe faaliyet gösteren büyük bir firmanın, ayrılma eğilimi gösteren müşterilerini belirleyerek; bu müşterilere özel pazarlama stratejileri geliştirilmesini hedeflemektedir. Ayrılacak müşteri profilini belirlemek için Lojistik Regresyon Analizi ve sınıflandırma tekniklerinden Karar Ağaçları kullanılmış ve uygulamanın sonuçları sunulmuştur.

Anahtar Sözcükler: *Müşteri ayrılma analizi, veri madenciliği, lojistik regresyon, karar ağaçları, telekomünikasyon.*

1. Introduction

For many companies, finding reasons of loosing customers, measuring customer loyalty and regaining customer have become very important concepts. Companies organize various studies and campaigns to avoid loosing their customers rather than to obtain new ones.

The telecommunication sector acquires huge amount of data due to rapidly renewable technologies, the increase in the number of subscribers and with value added services. Uncontrolled and very fast expansion of this field cause increasing losses depending on fraud and technical difficulties. Therefore, the developments of new analysis methods have become a must.

[1] *tugbasim@istanbul.edu.tr (U.T. Şimşek Gürsoy)*

In Turkey, this matter has become a pioneer in various researches in telecommunication sector which suffers from huge losses of customers. Churn Analysis which is often used in all sectors, is one of the applications of Data Mining. By estimating customers who will most likely switch service provider, organizations can create campaigns that aim to increase customer loyalty and develop marketing strategies to have higher customer retention.

The purpose of this study is to determine the reasons why the telecom company looses its customers. Like determining reasons, figuring out what types of customers are lost, is researched, as well.

2. Related Literature

Firms in telecommunication sector have detailed call records. These firms can segment their customers by using call records for developing price and promotion strategies [1].

By using Data Mining techniques, the subscribers who are intended not to make any payments, can be detected from before. And also, financial losses can be prevented. For this type of analysis, Deviation Determination method is applied. According to usage patterns subscribers are divided into specific clusters. The ones showing inconsistent features are determined and will be reviewed. By using Data Mining Techniques, International Roaming Agreements can also be optimized.

Data mining algorithms and knowledge discovery framework have been successfully applied in a number of application domains including commerce, astronomy, geological survey, security, and telecommunications [2]. Some of these are explained below.

In their study Ren, Zheng and Wu [3] presented a clustering method based on genetic algorithm for telecommunication customer subdivision. First, the features of telecommunication customers (such as the calling behavior and consuming behavior) are extracted. Then, the similarities between the multidimensional feature vectors of telecommunication customers are computed and mapped as the distance between samples on a two-dimensional plane. Finally, the distances are adjusted to approximate the similarities gradually by genetic algorithm.

Churn prediction and management have become of great concern to the mobile operators. Mobile operators wish to retain their subscribers and satisfy their needs. Hence, they need to predict the possible churners and then utilize the limited resources to retain those customers. In response to the difficulty of churner prediction, Chang's [4] study applies data mining techniques to build a model for churner prediction. Through an analysis result from a big Taiwan telecom provider, the results indicated that the proposed approach has pretty good prediction accuracy by using customer demography, billing information, call detail records, and service changed log to build churn prediction mode by using Artificial Neural Networks.

In a large software system knowing which files are most likely to be fault-prone is valuable information for project managers. They can use such information in prioritizing software testing and allocating resources accordingly. Turhan, Koçak and Bener [5] analyzed twenty five projects of a large telecommunication system in their study. To predict defect proneness of modules they trained models on publicly available Nasa MDP data. In their experiments they used Static Call Graph Based Ranking as well as Nearest Neighbor Sampling for constructing method level defect predictors.

There are many type of fraud in telecommunication sector. Hilas's [6] paper deals with the detection of fraudulent telecom activity inside large organizations' premises. Focus is given on superimposed fraud detection. The problem is attacked via the construction of an expert system which incorporates both the network administrator's expert knowledge and knowledge derived from the application of data mining techniques on real world data. In his paper, 22000 phone records which were made on 5541 days, were analyzed with the help of Expert Systems.

Daskalaki, et al., [7] reported on the findings of a research project that had the objective to build a decision support system to handle customer insolvency for a large telecommunication company. The main point of this study is to predict churner customers by using Decision Trees and Neural Networks.

Wei and Chiu [8] proposed design, and experimentally evaluate a churn-prediction technique that predicts churning from subscriber contractual information and call pattern changes extracted from call details. This proposed technique is capable of identifying potential churners at the contract level for a specific prediction time-period. The largest telecom company in Taiwan that has 21 million subscribers, were selected for application. They expressed that, the more call records they have, the more accurate results they can have from Churn analysis.

It is true for today that all of telecommunication companies use data mining in Turkey.

Vodafone which bought Telsim applies data mining for sales, marketing, financial management, future prediction, and for many different needs. Vodafone detects peak hours by using its databases and makes more workforces ready to avoid any disruption in communication. Also, Vodafone determines average of prepaid minutes purchased and finds subscribers who will likely churn [9].

Turkcell which obtained customer information through business intelligence and data mining techniques offers new tariffs and develops campaigns for existing customers. Turkcell also has been developing programs that increase customer loyalty. Credit Rating Project which is initiated to get new information about different payment behavior of customers is being set up [10]. Turkcell does not only increase customer satisfaction, but also benefits from reduced costs and risks.

In Turkey the system keeping the phone number while switching provider is established at the end of 2008. Avea which gained total of 463000 customers from rivals, take 335000 and 128000 customers from Turkcell and Vodafone, respectively [11]. In this period, Avea is the only operator that gained subscribers from rival operators.

3. Customer Churn

If a customer terminates a membership agreement with one company and become a customer of another competitor, this customer is called as lost customer or Churn customer.

Customer loss is very closely related with customer loyalty. Today’s economic trend dictates that price cuts are not the only way to build customer loyalty. Accordingly, adding new value added services to the products has become an industry norm to have loyal customer. The main goal of customer lost study is to figure out a customer who will likely be lost and is to calculate cost of obtaining those customers back again. During the analysis, the most important point is the definition of the churner customer. In some cases, to make a definition is very difficult. A credit card customer, for instance, can easily start using another bank’s credit card without cancelling credit card of current bank. In this specific case, a decrease in spending can be taken into consideration to understand the customer's loss. Customer’s loss is a major problem for companies which are likely to loose their customers easily. Banks, insurance and telecommunication companies can be given as examples.

For companies, the cost of acquiring new customers is increasing day by day. Therefore, a new era has begun in marketing industry. Instead of organizing campaigns to win new customers, companies are searching different variety of programs to emphasis on customer satisfaction, to increase customer-based earnings and to have higher customer loyalty. The only method to achieve those goals is preventing customer churn before it happens. At this point, customer churn modeling has created an important competitive advantage and a new workspace. A good modeling reveals which customer is close to churn and which is loyal.

With the development in database systems and the variability of customer behavior, an extraordinary increase in the size of the data has occurred. This causes to extract previously unknown information and relationships in huge amount of data. This information requires applying different techniques according to the structure of the data sets to be analyzed. The results of the analysis are used to plan a comprehensive promotional campaigns and new strategies.

4. Churn Analysis in Telecommunication Sector

Because of globalization and market conditions, Customer Relationship Management is becoming more and more important for all sectors.

Telecommunication sector provides services like rapidly growing local and intercity calls, and other communications services such as voice, fax, e-mail and other data traffic. Telecommunication market is rapidly developing and becoming competitive in many countries due to regulations, new computer and communication technologies. In this situation, data mining is required for understanding the business needs, defining the telecommunications model, using sources effectively and improving service quality.

Telecommunication industry is now a mature market and aware of the importance of Customer Relationship Management. Because of having mature market, developments on following fields are achieved.

- **Cross Selling and up-selling:** to maximize profits from existing customers.
- **Retaining and up-selling:** to retain profitable customers or get rid of inappropriate customers to the company profile.
- **Poaching:** to poach new Customers from rival companies.

Obtaining new customers are more expensive than retaining existing customers. It is for that, telecommunication companies realize that to keep existing customers is getting more important and agrees that churn analysis is one of the important data mining application areas.

Churn Analysis is applied to research why customers switch service provider. In a Churn analysis applications, the first thing is to access to the customer data. Then, factors are classified to decide which factor or factors affect customer churn decision. After determining which customers are likely to churn, different and specific marketing and retention strategies can be applied to the target customers, in a defined time period. Churn Analysis is not just applied in marketing; it is also applied in customer service, sales and finance applications. This departments need to identify what the possible results of churn are, how much financial impact the company has, how sales and customer service area are affected by churn.

Customers may leave the organization because of different reasons. That’s why different definitions can be done about the customer churn. Churner customer means “Customers who leave the company because of some reasons”. Churner customer can be categorized according to the party making the first move.

If the customer begins the first movement, this is called "**voluntary churn**". The termination of contract, the phone device changes, the service quality, the competition, the technological changes, the regulatory changes can be listed as reasons for loosing customers. If the organization is the starter; this is called as "**non-voluntary churn**". In this case, for some reason the company may decide to terminate their services to the customer. Unpaid bills for several months or not to load prepaid minutes are the most known reasons why service providers terminate contracts.

Based on a research, the percentage of Carrying Mobile Number (CMN) in developed countries is 27%.

The causes and rates of churn criteria are shown in Table 1. All of the tables and figures are given in Appendix.

The first reason of CMN is price. Churning to cheap operator is realized as normal. Churn rate resulting from price is 48.3%. This rate forces service providers to establish lower charge rates for different plans to sustain market share. The second reason for CMN is operator's width of service area. This rate is 19.4%. Another factor of CMN is the dissatisfaction of customers about service quality. The rate is 9.5%. The ads also affect the CMN. 4.1% of customers who change operators are affected by advertising. The percentage of the customers who will churn just for curiosity is 5.7%. The rate of switching just because of curiosity is higher than the rate of switching because of ads. This is a very interesting output and it is worth to make research on this issue.

5. Application of Churn Analysis

In this study, a customer data of a major Telecommunication Firm in Turkey is analyzed to estimate the churn probability. SPSS Clementine program is used for data analysis.

5.1. Business Understanding

This initial phase of data mining focuses on understanding the objectives of the project and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives [12]. Churn Analysis uses the data period in which customers are still with company, and focuses on customer retention. Customer retention consists of "Identifying which customers are likely to Churn, determining which customers should retain and developing strategies to retain profitable customers".

The main thing in retention process is identifying Churn ratio which is a very meaningful and vital determination for many companies. Determination of Churn ratio indicators is also very important. By using those indicators, firms can make prediction on future behavior of new customers and can develop new strategies much before customers start to think about churn. Thus, it is vital to build a very successful and accurate Churn model during the retention studies.

The aim of this study is to determine customers who churn from the services and products of the company that telecommunication firm has to keep the customers happy. Logistic Regression and Decision Trees are two important techniques that are applied.

5.2. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step [12].

Data which is collected from a Telecommunication Company to get analyzed, involves usage details of customers from 13.02.2009 till 08.06.2009. For the mentioned period, the number of records is 1000.

5.3. Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and elimination of data for modeling tools [12].

In this study, while one variable is being examined, three sub-variables are derived. These three sub-variables are average, standard deviation and trend. In dataset, for example, those three sub-variables for CALLOUT are determined as;

- AVRG_CALLOUT_5_MIN_ABOVE (AVRG_CALLOUT_5_DK_USTU)
- STD_CALLOUT_5_MIN_ABOVE (STD_CALLOUT_5_DK_USTU)
- CALLOUT_5_MIN_ABOVE_S (CALLOUT_5_DK_USTU_S)[*]

In Table 2, some variables and their explanations are listed.

According to the information received from the company, IND_PAK_FLAG variable is coded as True and False. ‘‘True’’ stands for non-discount situation and ‘‘False’’ stands for discount situation.

There is no incorrect, incomplete or duplicated record in dataset. Therefore, the dataset is ready for the selection stage.

Data selection process differs according to the model that will be built. For this applied Churn Analysis, the priority is to identify the dependent variable. For this specific application, the dependent variable is determined as ‘‘TARGET_FLAG’’. For the dependent variable, “1” represents customers who churn, and “0” in contrast, represents customers who keep current service.

5.4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Since some techniques have specific requirements on the form of data, stepping back to the data preparation phase is often needed [12].

For our study, since the dependent variable has two outcomes such as “0” and “1”, Logistic Regression which is appropriate for those kind of analysis, is applied. Before modeling stage, the distribution of Churn is examined. The distribution can be seen in Figure 1. Based on this figure, 74.8% of subscribers keep their service with firm, 25.2% of subscribers, in contrast, churn from company. If the data set is analyzed by using these ratios, outcomes will be biased. In order to have unbiased results, the distribution should be balanced.

Figure 2 shows balanced distribution. Based on this figure 49.97% of subscribers keep their service with firm, 50.03% of subscribers, in contrast, churn from company.

After balancing target variable, data set has become ready for modeling. The next step is to eliminate variables which have no effect on target variable (Figure 3).

All variables, after this elimination, can be considered as significant. It is very important to examine correlation among these variables to figure out whether any of two variables are correlated. By using statistics node in SPSS Clementine, variables which have effect on target variable and are not correlated are chosen. After data preparation stage, data

[*] Dataset received from Telecommunication Company is in Turkish. Therefore the original variables that are in Turkish and translated variables are shown together.

set become appropriate for Logistic Regression Analysis. The built model can be seen in Figure 4.

Significance levels are often used by analysts in reporting test results. Statistical tests should have significance levels of 0.05 or 0.01 [13]. We assumed significance level as 5% for this study.

β parameter, Wald statistics, degrees of freedom, significance levels and Odds statistics can be seen in Table 3.

Odds ratio is one category divided by the other. Odds (Exp(B)), represents the ratio which is calculated by using probability of staying with current operator and the probability of churner subscribers [14].

The Wald test is a parametric statistical test with a great variety of uses. Whenever a relationship within or between data items can be expressed as a statistical model with parameters to be estimated from a sample, the Wald test can be used to test the true value of the parameter based on the sample estimate. A Wald test is used to test the statistical significance of each coefficient (β) in the model. The statistics are used to test the null hypothesis that is H_0 = 0, where “0” is a vector with all entries equal to “0” [15]. The Wald statistics rejects the null hypothesis when the significance level is higher than 5%. For example; the significance level of AVRG_CALLOUT_GSM_DNUM variable is 0.889. It is higher than 5%. Hence H_0 hypothesis is rejected.

All of the variables which have higher significance level than 5%, are eliminated from the model. Then, the logistic regression model is applied again and the results are shown in Table 4. As it can seen in Table 4, all of the variables are significant, except CALLIN_5_MIN_ABOVE_S. This variable is eliminated and the model is applied one more time. The results are shown in Table 5. After a couple of eliminations, Table 5 which shows all significant variables, are obtained.

Decision Trees Analysis is also used for checking the Logistic Regression Analysis results. Figure 5 shows the model. As a result of different tries, the most appropriate decision tree algorithm, C&RT (Classification and Regression Trees) has been selected.

Figure 6 indicates that 50% percent of subscribers are staying with current operators and 50% of them is likely to churn. The number of non-churner subscribers is 748.

If the subscribers have discount package (IND_PAK_FLAG=0);

If the average of local and long distance calls is higher than 218, 42% of the subscribers will likely to Churn. If average of incoming calls from the same operator is equal to or lower than 4399, 72.9% of those subscribers will likely to Churn.

If the average of local and long distance calls is lower than 218, 62.3% of the subscribers will likely to Churn. If those subscribers don’t have prepaid plan, 67.9% will stay with service provider. This indicates that 53.7% of subscribers who have prepaid plan will switch their service provider.

If time of receiving call from the same provider is less than 4085, 63.3% of those subscribers will churn.

And finally, if the standard deviation of calls which is received from other GSM providers, is higher than 109, 74.4% of those subscribers will churn.

If the subscribers do not have discount package (IND_PAK_FLAG=1);

It is indicated that 75% of the subscribers are likely to churn.

If membership period is less than 1744, 80% of these subscribers tend to change the operator.

If membership period is less than 1744 and standard deviation of incoming calls from other operators is less than 121, 69.6% of these subscribers will stay with the current operator. If membership period is less than 1744 but standard deviation of incoming calls from other operators is more than 121, 87.7% of these subscribers tend to change the operator.

If membership period is less than 1744, standard deviation of incoming calls from other operators is less than 121 and membership period is less than 202, it is observed that 93.1% of those subscribers will churn.

If membership period is more than 1744, 59% of those customers will stay.

And finally, it is found that if the incoming calls from same operator are more than 2599, 85.7% of those customers will stay with their operator.

5.5. Evaluation

At this stage of a project you have build a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, review the steps executed to construct the model, and to be certain that it properly achieves the business objectives. A key objective is to determine whether there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached [12].

Table 6 can be used to evaluate the results of Logistic Regression model. According to significance level in Table 6 (less than 5%), the logistic regression model is significant.

The Nagelkerke R^2 statistic is a value similar to the variance in multiple regression [16]. In this case the Nagelkerke R^2 value is 0.274. Therefore, the model explained 27.4% of the variation in the dependent variable [17]. (Table 7)

Cox and Snell R^2 is an attempt to imitate the interpretation of multiple R-square based on the likelihood, but its maximum can be (and usually is) less than 1. The ratio of the likelihoods reflects the improvement of the full model over the intercept model (the lower the ratio, the greater the improvement). For this study the ratio is 0.205. It can be said that this rate is not high enough, therefore different variables are needed for data set and number of sampling must be more.

In Table 8, it can be seen that the classification rates for model built with IND_PK_FLAG, PREPAID_FLAG, AOL_DAYS, AVRG_CALLOUT_PSTN_DUR, AVRG_CALLIN_ONNET_DUR, STD_CALLIN_GSM_DUR, AVG_SMSMO_TRCELL_NUM, AVG_SMSMO_VF_NUM variables, is 70.1%.

In Logistic Regression Analysis, the accuracy of correctly predicting non-churners is 74.3% and the accuracy of predicting the churner subscribers correctly is 66%.

For evaluating C&RT Decision tree algorithm, the accuracy of the model is calculated and the result is 71.76% which is considered as very high. The probability of having wrong outcome is 28.24%. (Figure 7)

The gaining chart of the model can be seen in Figure 8. Based on this graph, if the company reaches 67% of the subscribers who has higher probability to churn, the company will actually be reaching 83.1% of the subscribers who will likely leave company.

5.6. Deployment

The creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be either as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the customer to fully understand which actions will be needed to carry out in order to actually make use of the created models [12].

The results obtained in this study can be used in marketing activities of the company. When the company organizes campaigns for its products and services, firm can give priority to the subscribers who are more likely to Churn, and can reduce the loss of customers.

6. Discussion

In this study, the data of a company which is operating in telecommunication sector is analyzed with data mining techniques with the aim of demonstrating models to predict churner customer behavior, improve customer relationship management, and develop various campaigns and marketing strategies for customer retention and loyalty. After eliminating non-related data and preparing stages, Logistic Regression Analysis and C&RT Decision Tree are applied for determining the reasons for customer churn.

The telecommunication sector, in which customers change providers rapidly, is very dynamic. If the profile of churner customers can be identified, specific campaigns can be created to keep the target groups.

The analysis shows that subscribers, who do not have a discounted package, have very high tendency to churn. So, various attractive packages should be created to fulfill different calling behaviors. Also, subscribers who don’t belong to any discount plan, should be informed about various plans. Thus, creating different marketing strategies and different plans that fulfill various customers’ profiles can help firms to keep customers happy and can reduce the number of churned customers.

The other high effective factor in churn analysis is the incoming local and long distance calls. Packages that are developed based on inner-city and long distance call records, can resolve concerns of subscribers.

It is showed that being under contract and receiving calls from subscribers, who are in the same service provider, are the reasons that increase commitment and loyalty to firm. Also, it is found that a decrease in the percentage of receiving calls through the same operator increases tendency of churn. Based on these outcomes, it can be said that the firm can promote switching from prepaid plans to under contract or postpaid plans. Moreover, to increase incoming calls from subscribers who are with the same operator discounted rates can be applied for in service calls.

When standard deviation of calls from other GSM operators increases, the churn rate increases too. In this case, making agreements with other GSM operators to provide the same or lower rates for between providers’ calls can be a wise maneuver. Also, plans such as one rate to all providers can be created and promoted. Moreover, subscribers should be informed about other services such as SMS, 3G, internet and so forth. Various plans that include some or all of those services should be promoted.

Associated with the information obtained from the analysis and based on presented recommendations, it can be promoted that the relationship with existing subscribers can be developed, the demand can be increased for company's different services, the churn rates can be minimized and the business profit margin can be increased.

References

[1] C. Rygielski, J.C. Wang, D.C. Yen, Data Mining Techniques for Customer Relationship Management. *Technology in Society*. 24(4), 483-502 (2002).

[2] P. Fule, Exploratory Medical Knowledge Discovery: Experiences and Issues. *ACM SIGKDD Explorations Newsletter*. 5(1), 94-99 (2003).

[3] H. Ren, Y. Zheng, Y. Wu, Clustering Analysis of Telecommunication Customers. *The Journal of China Universities of Post and Telecommunications*. 16(2), 114-116 (2009).

[4] Y.T. Chang, Applying Data Mining to Telecom Churn Management. *International Journal of Reviews in Computing.* 69-77, (2009).

[5] B.Turhan, G. Koçak, A. Bener, Data Mining Source Code for Locating Software Bugs: A Case Study in Telecommunication Industry. *Expert Systems with Applications*. 36, 9986- 9990 (2009).

[6] C.S. Hilas, Designing an Expert System for Fraud Detection in Private Telecommunications Networks. *Expert Systems with Applications.* 36, 11559-69 (2009).

[7] S. Daskalaki, et al., Data Mining for Decision Support on Customer Insolvency in Telecommunications Business. *European Journal of Operational Research*. 145, 239-255 (2003).

[8] C.P. Wei, I.T. Chiu, Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Systems with Applications*. 23, 103-112 (2002).

[9] N. Akkaş, *Kahin Şirketlerin Kehanetleri*. http://www.sas.com/offices/europe/turkey/news/basindasas/inthenews_new_010908.htm (2010), (Erişim: 15.02.2010).

[10] E. Acar, *Turkcell SAS Veri Madenciliği Çözümü İle Abonelikten Terk Oranını Yüzde 50 Azalttı*. http://www.sas.com/offices/europe/turkey/news/pressreleases/october2005/news02_131005.htm (2005), (Erişim: 08.01.2010).

[11] M. Çehreli, *Numara Taşınabilirliğinde 463000 Kişinin Avea'ya Geçtiği Açıklandı.* http://www.turk.internet.com/portal/yazigoster.php?yaziid=23087 (2009), (Erişim: 02.02.2010).

[12] *Cross Industry Standard Process for Data Mining*, www.crisp-dm.org (2010), (Erişim: 05.02.2010).

[13] M.L. Cohen, J.E. Rolph, D.L. Steffey, Statistics, *Testing and Defense Acquisitions: New Approaches and Methodological Improvements*. The National Academies Press, 1998, p.91.

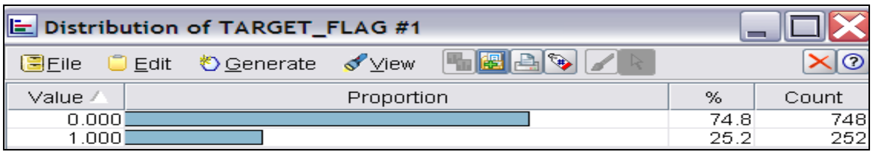
[14] J. Walker, Methodology Application: Logistic Regression Using The CODES Data. D*epartment of Transportation National Highway Traffic Safety Administration (NHTSA) and National Center for Statistics and Analysis.* 8 (1996).

[15] T. Fears, J. Benichou, M.H. Gail, A Reminder of The Fallibility of The Wald Statistics. *The American Statistician.* 50 (1996).

[16] M.J. Norusis, Straight Talk About Data Analysis and SPSS. *SPSS Professional Statistics. Chicago: SPSS, Inc.* (1997).

[17] M.F. Kraska, J.R. Larkins, Factors Affecting Master Sergeants' Completion of Community College of the Air Force AAS Degree Requirements. *Journal of Information Techonology Education*. 36 (3) (1999).

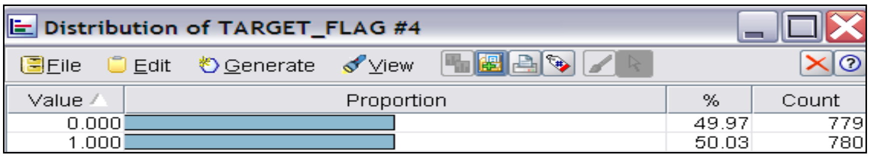
Appendix-1 Figures



Distribution of TARGET_FLAG #1

Value	Proportion	%	Count
0.000		74.8	748
1.000		25.2	252

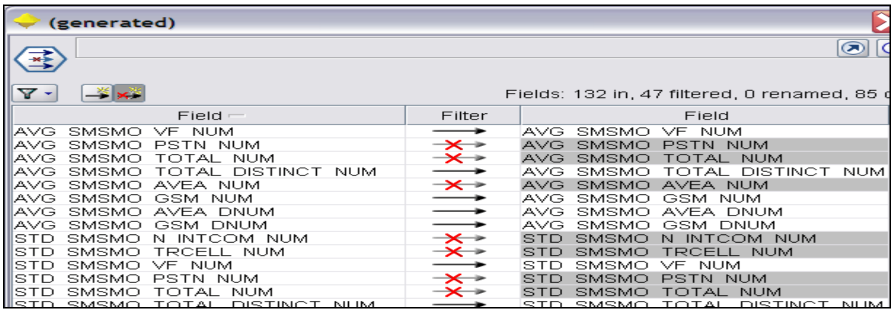
Figure 1 The Distribution of Dependent Variable in Data set



Distribution of TARGET_FLAG #4

Value	Proportion	%	Count
0.000		49.97	779
1.000		50.03	780

Figure 2 Dependent Variable in the Data Set Balanced



(generated)

Fields: 132 in, 47 filtered, 0 renamed, 85 c

Field	Filter	Field
AVG SMSMO VF NUM	→	AVG SMSMO VF NUM
AVG SMSMO PSTN NUM	✗→	AVG SMSMO PSTN NUM
AVG SMSMO TOTAL NUM	✗→	AVG SMSMO TOTAL NUM
AVG SMSMO TOTAL DISTINCT NUM	→	AVG SMSMO TOTAL DISTINCT NUM
AVG SMSMO AVEA NUM	✗→	AVG SMSMO AVEA NUM
AVG SMSMO GSM NUM	→	AVG SMSMO GSM NUM
AVG SMSMO AVEA DNUM	→	AVG SMSMO AVEA DNUM
AVG SMSMO GSM DNUM	→	AVG SMSMO GSM DNUM
STD SMSMO N INTCOM NUM	✗→	STD SMSMO N INTCOM NUM
STD SMSMO TRCELL NUM	✗→	STD SMSMO TRCELL NUM
STD SMSMO VF NUM	→	STD SMSMO VF NUM
STD SMSMO PSTN NUM	✗→	STD SMSMO PSTN NUM
STD SMSMO TOTAL NUM	✗→	STD SMSMO TOTAL NUM
STD SMSMO TOTAL DISTINCT NUM	→	STD SMSMO TOTAL DISTINCT NUM

Figure 3 Eliminated Variables by Filter Node

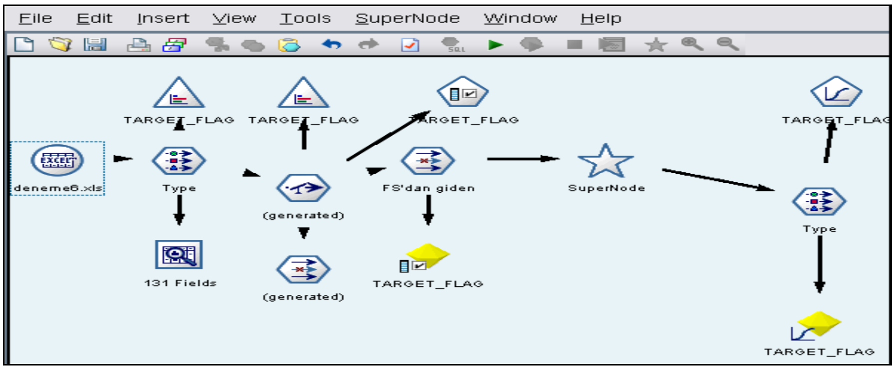


Figure 4 Logistic Regression Model

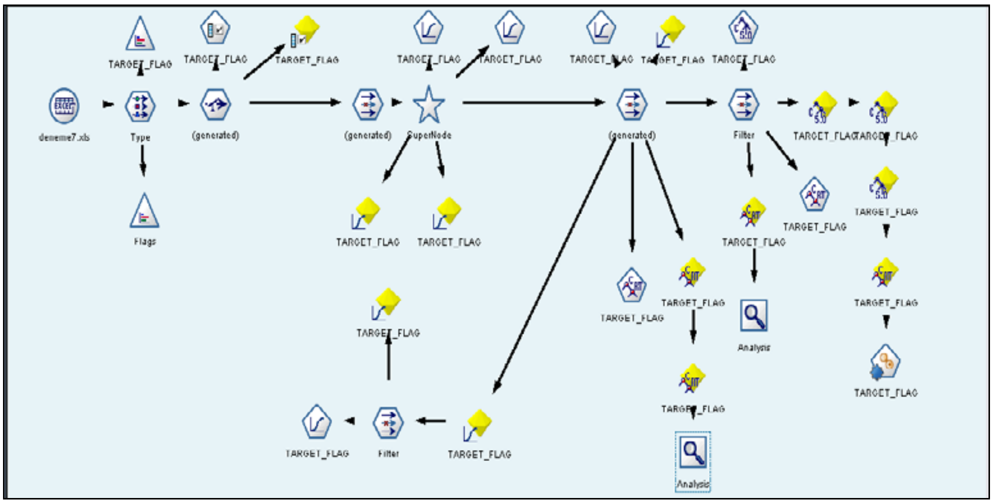


Figure 5 Decision Tree Model

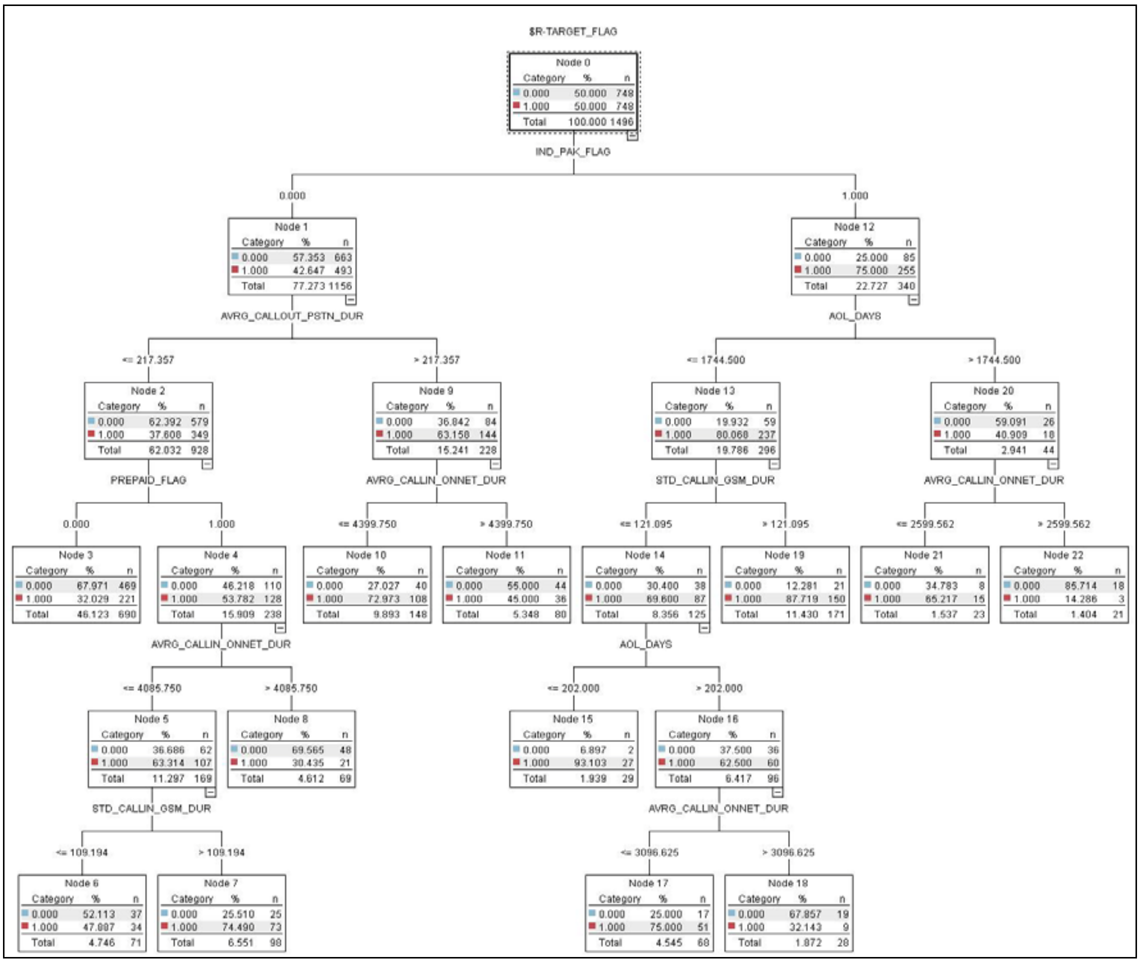
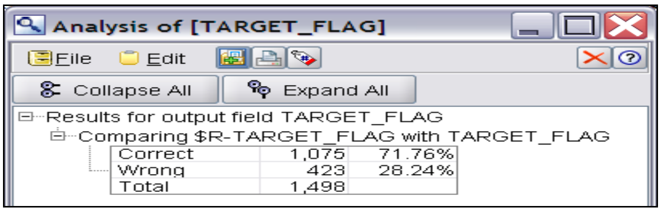


Figure 6 C&RT Decision Tree



Results for output field TARGET_FLAG

Comparing $R-TARGET_FLAG with TARGET_FLAG

Correct	1,075	71.76%
Wrong	423	28.24%
Total	1,498	

Figure 7 C&RT Decision Trees Analysis

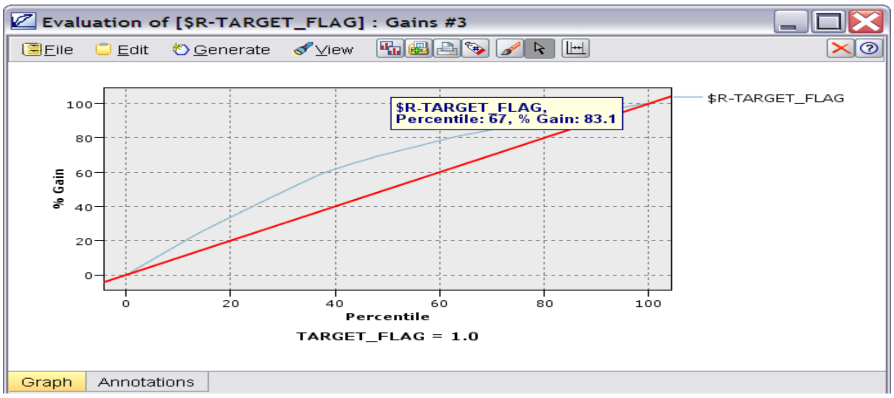


Figure 8 Gaining Chart of a Model

Appendix-2 Tables

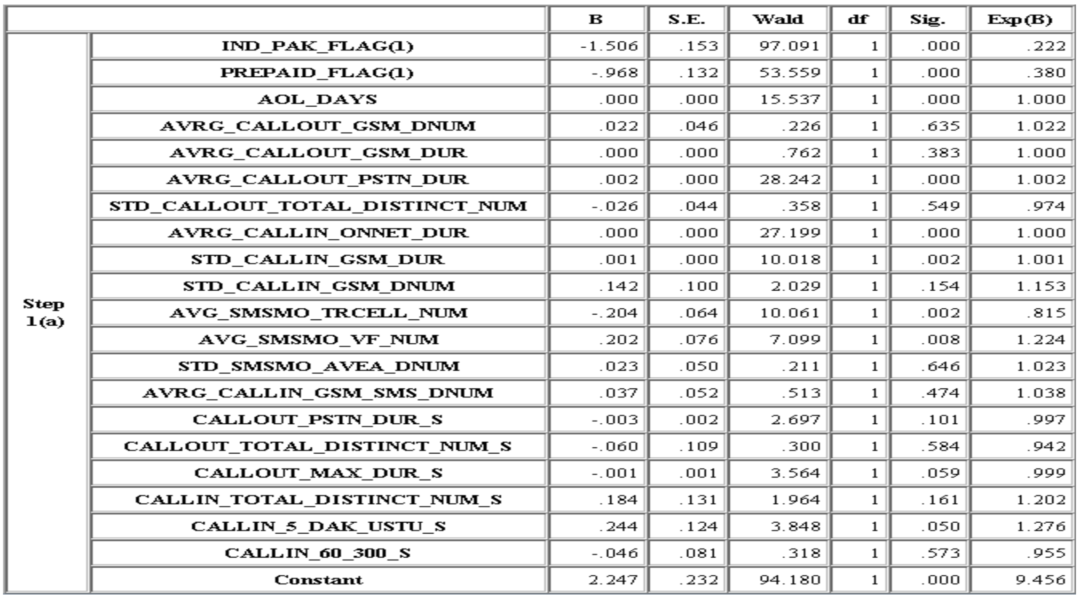
Tablo 1 The Causes of Churn [11]

Causes	Percent
Price	48.3
Width of Service Area	19.4
Service Quality	9.5
Advertisement	4.1
Pre-paid	13.0
Curiosity	5.7

Table 2 Variables and Explanations

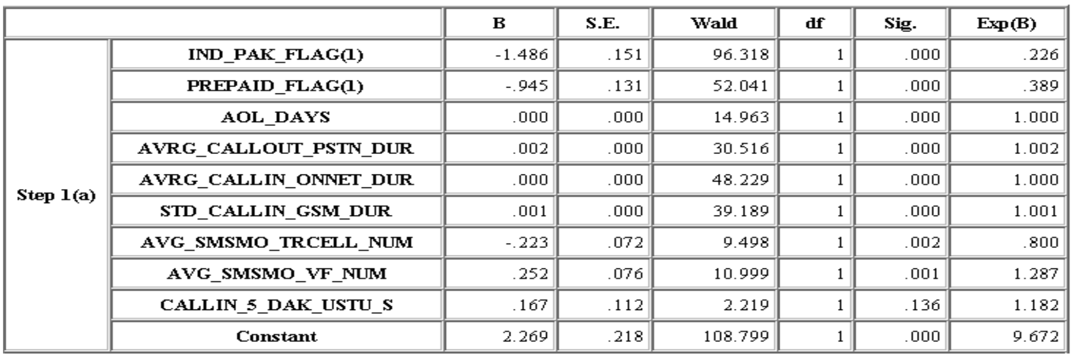
Variables	Explanation
CUSTOMER ID	Customer ID
TARGET_FLAG	Target churn flag
IND_PAK_FLAG	With or without discount package
PREPAID_FLAG	Prepaid/Postpaid
TEK_HAT_PK_FLAG	With or without second number
AOL_DAYS	Age of line (Days)
CURRENT_TARIFF_GROUP	Current tariff group
AVRG_CALLOUT_5_DAK_USTU	Average of calls more than 5 min.
AVRG_CALLOUT_60_300	Average of calls between 60-300 min.
AVRG_CALLOUT_FIRM_DNUM	Average number of callouts to same operator
AVRG_CALLOUT_FIRM_DUR	Average duration of callouts to same operator
AVRG_CALLOUT_GSM_DNUM	Average number of callouts to all GSM operators
AVRG_CALLOUT_GSM_DUR	Average duration of callouts to all GSM operators
AVRG_CALLOUT_INTCOM_DUR	Average of calls made to same operator and same package
AVRG_CALLOUT_MAX_DUR	Average of maximum calls
AVRG_CALLOUT_MIN_DUR	Average of minimum calls
AVRG_CALLOUT_N_INTCOM_DUR	Average of calls made to same operator and different package
AVRG_CALLOUT_PSTN_DUR	Average of local and long-distance calls
AVRG_CALLOUT_TOTAL_DISTINCT_NUM AVRG_CALLOUT_TOTAL_DUR	Average of total unique number calls Average of total calls
AVRG_CALLOUT_TOTAL_NUM	Average of total called numbers
STD_CALLOUT_1_DAK_ALTI	Standard deviation of calls less than 1 min.
STD_CALLOUT_5_DAK_USTU	Standard deviation of calls more than 5 min.
STD_CALLOUT_60_300	Standard deviation of calls between 60-300 min.
STD_CALLOUT_FIRM_DNUM	Standard deviation of callouts to same operator

Table 3 Logistic Regression Result -1-



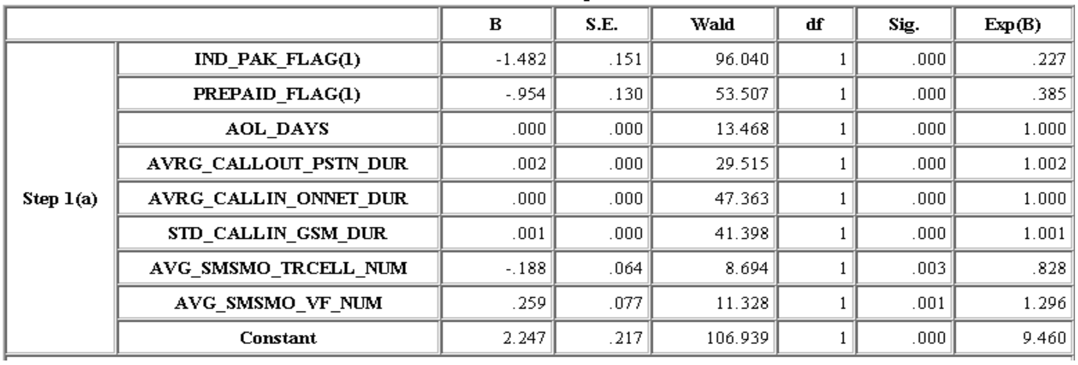
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	IND_PAK_FLAG(1)	-1.506	.153	97.091	1	.000	.222
	PREPAID_FLAG(1)	-.968	.132	53.559	1	.000	.380
	AOL_DAYS	.000	.000	15.537	1	.000	1.000
	AVRG_CALLOUT_GSM_DNUM	.022	.046	.226	1	.635	1.022
	AVRG_CALLOUT_GSM_DUR	.000	.000	.762	1	.383	1.000
	AVRG_CALLOUT_PSTN_DUR	.002	.000	28.242	1	.000	1.002
	STD_CALLOUT_TOTAL_DISTINCT_NUM	-.026	.044	.358	1	.549	.974
	AVRG_CALLIN_ONNET_DUR	.000	.000	27.199	1	.000	1.000
	STD_CALLIN_GSM_DUR	.001	.000	10.018	1	.002	1.001
	STD_CALLIN_GSM_DNUM	.142	.100	2.029	1	.154	1.153
	AVG_SMSMO_TRCELL_NUM	-.204	.064	10.061	1	.002	.815
	AVG_SMSMO_VF_NUM	.202	.076	7.099	1	.008	1.224
	STD_SMSMO_AVEA_DNUM	.023	.050	.211	1	.646	1.023
	AVRG_CALLIN_GSM_SMS_DNUM	.037	.052	.513	1	.474	1.038
	CALLOUT_PSTN_DUR_S	-.003	.002	2.697	1	.101	.997
	CALLOUT_TOTAL_DISTINCT_NUM_S	-.060	.109	.300	1	.584	.942
	CALLOUT_MAX_DUR_S	-.001	.001	3.564	1	.059	.999
	CALLIN_TOTAL_DISTINCT_NUM_S	.184	.131	1.964	1	.161	1.202
	CALLIN_5_DAK_USTU_S	.244	.124	3.848	1	.050	1.276
	CALLIN_60_300_S	-.046	.081	.318	1	.573	.955
	Constant	2.247	.232	94.180	1	.000	9.456

Table 4 Logistic Regression Result -2-



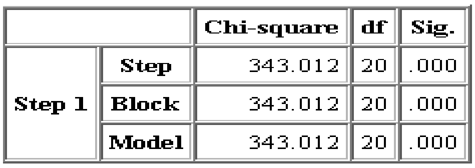
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	IND_PAK_FLAG(1)	-1.486	.151	96.318	1	.000	.226
	PREPAID_FLAG(1)	-.945	.131	52.041	1	.000	.389
	AOL_DAYS	.000	.000	14.963	1	.000	1.000
	AVRG_CALLOUT_PSTN_DUR	.002	.000	30.516	1	.000	1.002
	AVRG_CALLIN_ONNET_DUR	.000	.000	48.229	1	.000	1.000
	STD_CALLIN_GSM_DUR	.001	.000	39.189	1	.000	1.001
	AVG_SMSMO_TRCELL_NUM	-.223	.072	9.498	1	.002	.800
	AVG_SMSMO_VF_NUM	.252	.076	10.999	1	.001	1.287
	CALLIN_5_DAK_USTU_S	.167	.112	2.219	1	.136	1.182
	Constant	2.269	.218	108.799	1	.000	9.672

Table 5 Logistic Regression Result -3-



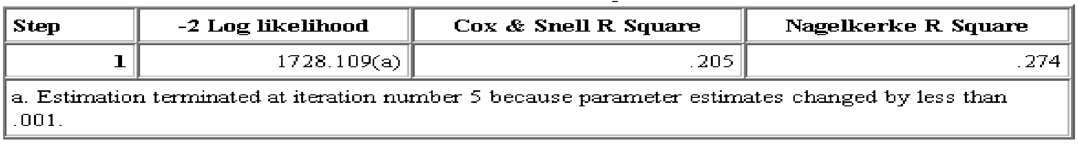
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	IND_PAK_FLAG(1)	-1.482	.151	96.040	1	.000	.227
	PREPAID_FLAG(1)	-.954	.130	53.507	1	.000	.385
	AOL_DAYS	.000	.000	13.468	1	.000	1.000
	AVRG_CALLOUT_PSTN_DUR	.002	.000	29.515	1	.000	1.002
	AVRG_CALLIN_ONNET_DUR	.000	.000	47.363	1	.000	1.000
	STD_CALLIN_GSM_DUR	.001	.000	41.398	1	.000	1.001
	AVG_SMSMO_TRCELL_NUM	-.188	.064	8.694	1	.003	.828
	AVG_SMSMO_VF_NUM	.259	.077	11.328	1	.001	1.296
	Constant	2.247	.217	106.939	1	.000	9.460

Table 6 Significance Test



		Chi-square	df	Sig.
Step 1	Step	343.012	20	.000
	Block	343.012	20	.000
	Model	343.012	20	.000

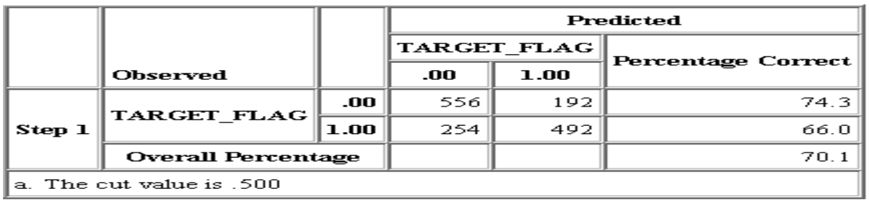
Table 7 Model Summary



Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1728.109(a)	.205	.274

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Table 8 Classification Rates



	Observed		Predicted		
			TARGET_FLAG		Percentage Correct
			.00	1.00	
Step 1	TARGET_FLAG	.00	556	192	74.3
		1.00	254	492	66.0
	Overall Percentage				70.1

a. The cut value is .500