➔ ***Regular Research Paper – NS***

Comparing to Techniques Used in Customer Churn Analysis

Ozer Celik
Department of Mathematics-Computer, Eskisehir Osmangazi University Faculty of Science, Eskisehir, Turkey
ozer@ogu.edu.tr

Usame O. Osmanoglu
Department of Biostatistics, Eskisehir Osmangazi University Faculty of Medicine, Eskisehir, Turkey
usa.osm@gmail.com

Abstract

In today's competitive conditions, the importance of minimizing costs is increasing day by day. As a result of the researches, it has been determined that the cost of attracting new customers is 10 times more than the cost of holding existing customers. This increases the importance of customer churn analysis, too. In this review, machine learning algorithms such as artificial neural networks (ANN), decision trees, support vector machines (SVM), naive bayes, k-nn and extreme gradient boosting (XGBoost) and customer churn analysis, Cox proportional hazard model and deep learning techniques. In addition, customer churn analysis studies conducted in various sectors using these techniques were examined. When the customer churn analysis studies are examined, it is seen that more complex systems can get modeling and reached higher success rates with deep learning technique. However, because it is a new technique and could not give stable results, machine learning algorithms, which are the closest alternative, are thought to be able to useful in estimating time-related events such as customer churn. The Cox regression model was found to be successful in estimating the independent variables affecting the time variable, the rate of life expectancy and the groups under risk. It is seen that deep learning techniques give better results in more complex structures.

Keywords: *Customer Churn Analysis, Machine Learning, Deep Learning, Cox Proportional Hazard.*

1. INTRODUCTION

1.1. What is Customer Churn?

The customer churn analysis can be defined as analytical work carried out on the possibility of a customer leaving a product or service. In its simplest definition, it means that customers are abandoned to choose the company because of competition [1]. The purpose is to identify this situation before leaving the customer's product or service, and then to carry out some preventive actions.

1.2. Importance of Churn Analysis

It is particularly important in the calculations of a business in sectors such as insurance, telecommunications or banking, which is a subscription-based income model. According to researchers, winning new customers in today's competitive conditions is up to 10 times costlier than retaining existing customers [2]. It is an analysis method used in areas such as determining profiles of existing customers, analyzing customer escapes and estimating customer escape [3].

Moreover, the value of these enterprises is directly proportional to the number of active customers. Therefore, many parameters such as the costs, profitability, size, investment capacity, cash flow of the enterprises depends on the number of customers and therefore the loyalty of the customers. In addition, research has shown that the profitability of long-term customers is higher. To understand this profitability, methods such as customer lifetime value (CLV/CLTV) are applied [4].

1.3. History

In the 1940s, scientists used ball and fire to explain the decision-making mechanism of human based on their research into the electrical collisions of neurons. Thus, artificial intelligence studies began in the 1950s [5]. In these years, Alan Turing carried out the Turing Test in order to test a machine's ability to imitate the human. The main purpose of the Turing Test is to measure the ability of the machine to communicate with people during a conversation. If the machine shows a less bad performance than the people, it is successful and pass the test. In 1956, the term "artificial intelligence" was used for the first time in a summer school organized by John McCarthy of the Massachusetts Institute of Technology and Allen Newell and Herbert Simon at Carnegie-Mellon University. In 1959, after Arthur Samuel produced a checkers programme, the machine learning was initiated. From these years until the 1980s, studies such as abstract reasoning and information-based systems were carried out and this process was called winter of artificial intelligence. In the 1990s, with the development of game technologies, artificial intelligence and machine learning activities increased rapidly. Today, artificial intelligence and machine learning is still used in many research and study fields [6].

Classical methods of analysis are insufficient in medical research. The first reason for this is that the evaluation of the research should be done before all the patients die or the emerge of the result that examined. Otherwise, it may take years to understand which treatment method is better and factors affecting the disease. The second reason is that the treatments applied to patients do not start at the same time. In these types of studies, Life analysis methods give more appropriate results [7]. Studies on the statistical analysis of failure or death time of the human and surrounding units started with the help of life table. These studies were then developed and named as failure model or hazard model. These models were used to analyze to time of the occurrence or observation of any well-defined event. Life analysis studies, which began in the 20th century, have shown great advances during the second half of this century. The most effective developments in this field;

- “Kaplan-Meier Method” [8], used for estimation of survival function,
- “Test Log-rank Test Statistics” used to compare two survival distributions [9],
- “Cox Regression Model” [10], used to measure the effects of explanatory variables on survival time,
- To reveal the small and large sample characteristics, survival analysis statistics have initiated with the "Counting Process with the Martingale Theory" which provided a unified structure, and it has come up to date [11].

Deep Learning was discovered by influencing the neurons depths of the brain. Researchers have endeavored to educate the multi-layered deep neural networks for years. Until 2006, studies conducted were able to train two- or three-layer neural networks, but studies on more layers failed. In 1960, the first generation of neural networks was discovered by Frank Rosenblatt. This structure which used a hand-made feature layer to identify objects by combining all the features and by finding the weight vector was called Perceptron. In 1985, Geoffrey Hinton replaced the fixed and single feature layer with several hidden layers using the Perceptron basis, thus creating the second-generation neural network. In 1995, Vladimir Vapnik discovered SVM (Support Vector Machines) with his colleague. The SVM was focused on statistical learning cores, and in contrast to Perceptron, fixed property layer was processing the single layer not directly, but after converting the entered single layer into multi-dimensional space. Although SVM has a simple structure, the learning is quick and easy. Although SVM solved many problems in

artificial intelligence, it has significant shortcomings and is a superficial architecture. In 2006, Geoffrey Hinton and Ruslan Salakhutdinov showed how could be effectively trained multilayer fed neural networks in an article they published. In this study, named "Deep Belief Network", showed how the multi-layered deep structures work and how to complete the completely unfinished features on their own. This artificial neural network was called Deep Learning [12].

2. THE STAGES OF CUSTOMER CHURN ANALYSIS

2.1. Examination of Data Set

Data set is a two-dimensional matrix with observation units take part in on the rows and variables take part in on the columns. The intersection of rows and columns is called a cell. The values of the observations (measurement results) are entered as numbers or symbols in each cell. If there is no number or symbol in the cell, it is called missing data.

The examination of the data set starts with checking the accuracy of the observation values first. If the data is not correct, it will cause serious problems such as wrong of the success rate of the model. Min-max control of variables and illogical situations are examined. For example, the age variable cannot be negative. After the observation values are determined to be correct, the missing data is eliminated because some mathematical operations cannot be performed. Also, since the variables should not be related to each other, only one of the associated variables can be included in the model, otherwise there is a multiple connection problem that reduces the reliability of the model.

The research field such as insurance company is selected for churn analysis. The purpose of the analysis is to establish a model that can predict whether customers will leave the service they receive and predict the possibility of leaving by time. For this model, it is necessary to create the correct data set. In previous studies, it shown to be included in the model demographic variables such as age and gender, as well as other variables related to the sector, which are suggested by experts, and thought to have an effect on churn. In addition, the data set problems should be taken into account when preparing the data set for the model creation by defining the missing values and excess values [13].

2.2. Establishing the Model

For the model to be balanced, stratified sampling is applied according to a variable such as age a way that the numbers of people in the sub-groups of the churn variable (churn = 1 and churn= 0) is equal. Binary logistic regression analysis is applied to dataset with variables likely to affect churn by using statistical software programs such as SPSS. During this process, the churn variable that can only have a value of 1 or 0 is defined as the dependent variable and the other variables is as the independent variable. In some software programs, categorical variables are self-encoded, and in some others, it is necessary to encode manually. As a result of the analysis, the variables affecting the churn are determined. The performance of the model with these variables can be tested by Machine Learning Algorithms, Cox Proportional Hazard Method and Deep Learning Methods.

2.3. Techniques Used in Customer Churn Analysis

2.3.1. Machine Learning

Learning has been described by Simon as the process of improving behavior through the discovery of new information over time. The learning is called Machine learning when perform by a machine. The concept of improvement is the status of finding the best solution for future problems by gaining experience from the existing examples in the process of machine learning [14]. With the development of information technologies over time, the concept of big data has emerged. The concept of big data is defined as very large and raw data sets that limitless and continue to accumulate, which cannot be solved by traditional databases methods [15].

The operations performed on the computer using the algorithm are performed according to a certain order without any margin of error. However, unlike the commands created to obtain the output from the data entered in this way, there are also cases where the decision-making process takes place based on the sample data already available. In such cases, computers can make the wrong decisions such as mistakes that people can make in the decision-making process. In other words, machine learning is to gain a learning ability like human brain to computer by taking advantage of data and experience [16].

The primary aim of machine learning is to develop models that can train to develop themselves and by detecting complex patterns and to create models to solve new problems based on historical data [17].

Machine learning and data-driven approaches are becoming very important in many areas. For example, smart spam classifiers protect our e-mails by learning from large amounts of spam data and user feedback. Ad systems learn to match the right ads with the right content; fraud detection systems protect banks from malicious attackers; Anomaly event detection systems help experimental physicists to find events that lead to new physics.

2.3.1.1. Customer Churn Analysis with Machine Learning

2.3.1.1.1. Training Data

After stratified sampling, the data set is divided into 70-30%, 80-20% as test-training data. The big one is used for training. Training is performed using various machine learning algorithms.

2.3.1.1.2. Test Data

The test data is the rest of the training data. At the end of the test process, the confusion matrix is used to calculate the success of the model. This matrix is a useful table that summarizes the predicted situation with the actual situation. Model success is calculated by the formula "(a+d)/(a+b+c+d)" (Table 1).

Table 1. Confusion Matrix

	Predict Situation		
Actual Situation	1	0	Ratio
1	True Positive (a)	False Negative (b)	(a+b)/(a+b+c+d)
0	False Positive (c)	True Negative (d)	(c+d)/(a+b+c+d)
Ratio			(a+d)/(a+b+c+d)

Huigevoort et al. estimated customer churn using data from Dutch health insurance companies with more than three million customers. They compared success rates by using logistic regression, decision tree, neural networks and support vector machine algorithms. They reported that the data set should be balanced to improve model performance. They observed that logistic regression gave the best result in 70-30% training set [13].

Tamaddoni et al. evaluated the performance of various parametric and nonparametric churn prediction techniques to define the optimal modeling approach based on the content by using empirical and simulated data from two online retailers. The results show that, in most cases (ie, the size of the sample, changing the buying frequencies and churn rates), the boosting technique, which is a nonparametric method, provides a superior estimate. Moreover, they reported that logistic regression was good in cases where churn was less. Finally, they reported that the parametrical probability models left behind other techniques when the number of customers was very small [18].

2.3.1.2. Machine Learning Algorithms

2.3.1.2.1. Artificial Neural Networks (ANN)

Artificial neural networks have been developed based on the human brain's biological neural networks and are an information processing system designed to perform the functions of these networks [19].

2.3.1.2.2. Decision Tree

Decision tree is the decision structure that performs learning from known data classes by inductive method. Decision tree is a learning algorithm that separates large amounts of data into small data groups using simple decision-making steps. As a result of each successful separation, the members in the result group are more like each other. Decision tree with descriptive and predictive features is one of the most preferred classification algorithms because of its easy to interpret, easy to integrate into databases and reliable [20].

2.3.1.2.3. Support Vector Machine (SVM)

The support vector machine is one of the supervised classification techniques laid down by Cortes and Rapnik (1995). SVM is the machine learning algorithm which makes prediction and generalization about new data by performing learning on data that unknown the distribution. The basic principle of the SVM is based on the presence of a hyperplane that best distinguishes the data of two classes. The support vector machine is divided into two according to the linear separation and nonlinear separation of the data set [21].

2.3.1.2.4. Naive Bayes

The Naive bayes classification is a classification using statistical methods for labeling data. Since it is easy to use, it is frequently preferred in classification problems. In general, it is aimed to calculate the probability values of the effects of each criterion in the Bayesian classification. Naive bayes calculates the conditional probability of the class to which the data belongs, in order to estimate the probability of a class with a data. Bayes theorem is used in this process.

2.3.1.2.5. Logistic Regression

Logistic regression is a method of classifying the relationship between multiple independent variables and dependent variables. Although it has usually been used in medical field in the past, it is an advanced regression method which has gained popularity in social sciences today. Logistic regression is a technique used as an alternative to this method due to the inadequacy of Least Squares Method (LSM) in a multivariate model with dependent and independent variable discrimination. In logistic regression analysis, the probability of the dependent variable with two values is predicted. In addition, the variables in the model are continuous. Because of this feature, it is a technique frequently used for classifying observations.

2.3.1.2.6. k-Nearest Neighbor (k-NN)

The k-nearest neighbor algorithm, which submitted by Fix and Hodges in 1951, are based on the logic that the data closest to each other belong to the same class. The main purpose is to classify the new incoming data by using the data previously classified. The data, which is unknown to which class it belongs to, are called test samples, the previously classified data are called learning samples. In the k-NN algorithm, the distance of the test sample from the learning samples is calculated, and then the k-learning sample closest to the test sample is selected. If the selected k samples have mostly belonged to which class; the class of the test sample is also determined as this class [22]).

2.3.1.2.7. Extreme Gradient Boosting (XGBoost)

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The name xgboost, though, actually refers to

the engineering goal to push the limit of computations resources for boosted tree algorithms. Therefore, many researchers use XGBoost. The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems [23].

2.3.2. Cox Regression

Survival analysis is concerned with a group of individuals or groups of individuals with a point event, often referred to as failure. Failure occurs after a certain time interval and this is called failure time. The time between a living organism or a lifeless object with a certain start time and death (failure) is called life time or failure time and is usually indicated by T. Examples of failure time include life expectancy of machine components, periods of unemployment in the economy, time of completing the task of the subject in a psychological experiment, and life expectancy of patients in a clinical trial [7]. Purposes of survival analysis; to obtain life expectancy estimates at different times, to estimate the distribution of life expectancy, to compare the life expectancy of different patient groups (Collett, 1994). In addition, intra-group risk rates of categorical variables can be calculated by Cox proportional risk model. For example; female customers are more likely to churn 50% more than male customers, middle-aged customers are more than 50% likely to churn according to customers in the elderly group.

2.3.2.1. Customer Churn Analysis with Cox Regression Model

Cox regression algorithm is run in pre-prepared test data and software programs such as Python. In output, in the proportional risk model generated by the training data is obtained survival probabilities decreasing from the first time each customer enters the test by using the test data. Then, with the expert opinion, a threshold value is selected, it thought that the customers will be churn in the time period corresponding to the threshold value, and then the churns are tried to be avoided with the necessary campaigns.

Jamal et al. analyzed the link between a time loss estimate and a heterogeneous hazard model for a South American television company [25].

Wong et al. used the Cox Regression technique to identify factors that led to customer escapism for a Canadian telecommunications company. A data set with demographic characteristics of 4896 people and various temporal variables was used. Contracted customers compare to customers who do not have a contract have a rate of 95.9% less probability to churn [26].

2.3.3. Deep Learning Technique

Deep Learning was discovered by influencing from the structure of the neurons depth of the brain. Learning technique which modelling multi-layered deep structures, and which be able to the self-complementing to unfinished features is called deep learning. Weibull Time To Event Recurrent Neural Network (WTTE-RNN) has been recently proposed for customer churn analysis. This technique is used to predict the next repetition time of the event in the case of repetitive situations. In addition, this model has been trained with log-likelihood-loss function for censored data which is frequently used in survival analysis. Weibull distribution is simple enough to prevent heterogeneity and overfitting. The estimated Weibull parameters can be used to estimate the expected value and duration of the next event. WTTE-RNN is defined using a general structure for censored data. This model can be easily expanded with other distributions and applied for multivariate estimation. In addition, in a study, it was determined that WTTE-RNN, Proportional Hazard Model and the Weibull Accelerated Failure time model had a special type [27].

2.3.3.1. Customer Churn Analysis with Deep Learning

A forecasting model is created by making likened to a specific distribution such as Weibull the distribution of the part of the training data. The test data is tested with this model. In this deep

learning model, the expected observation time of the next event is calculated using Weibull parameters. The WTTE-RNN deep learning technique is a technique derived from the Proportional Hazard model. However, this technique can be used in the case of repetitive events. Because such deep learning techniques are still in development, their reliability will increase with time.

Castanedo et al. applied deep learning networks and multilayered forward feed networks with different configurations, to predict customer loss by using four-layer high-tech architecture. In the study, it was recommended deep learning to estimate billions of call records from a corporate business intelligence system and a deduction in a prepaid mobile telecommunications network. The AUC achievement value of 77,9% was reached with the established model and a significantly better result was achieved compared to 73,2%, which is the best performance obtained by Random Forest machine learning technique [28].

Wangperawong et al. estimated customer churn in the telecommunications industry in Thailand by utilizing deep learning architecture. Calculated with the data set with 6 variables of 6 million customers. The results of simple learning techniques such as decision tree were compared with the results of deep learning technique. As a result of the study, AUC scores were calculated between 66.5% and 77.8%. They reported that the highest performance belongs to the deep learning architecture [29].

Martinsson in the case of discrete or continuous censored data, recurrent events, or time-varying common variables, time-event established a WTTE-RNN model for a sequential estimate [27].

Spanoudes et al. used the deep learning technique to predict churn through customer records in a company running subscription system. They reported that deep learning technique was not worse than the current technique of the company and that the model they established should be tested on more companies. With the help of the development of deep learning techniques, it is believed that higher achievements will be achieved [30].

3. CONCLUSION

When the customer churn analysis studies are examined, it is seen that more complex systems can get modeling and reached higher success rates with deep learning technique. However, because it is a new technique and could not give stable results, machine learning algorithms, which are the closest alternative, are thought to be able to useful in estimating time-related events such as customer churn. In order to increase the success, the data set should be balanced, sufficiently large, systematic error-free, and the explanations of the independent variables should be sufficient. The Cox regression model was found to be successful in estimating the independent variables affecting the time variable, the rate of life expectancy and the groups under risk. It is seen that deep learning techniques give better results in more complex structures. Furthermore, it is predicted that higher success rates will be achieved with improving deep learning techniques over time.

REFERENCES

[1] Nettleton, D. (2014). Commercial data mining: processing, analysis and modeling for predictive analytics projects. Elsevier.

[2] Kotler, P., & Keller, K. L. (2009). Manajemen pemasaran.

[3] Seker, S. E. (2016). Müşteri Kayıp Analizi (Customer Churn Analysis). YBS Ansiklopedi, 3(1), 26-29.

[4] Poel, V. D., Lariviére (2004). Customer Attrition Analysis For Financial Services Using Proportional Hazard Models. European Journal of Operational Research 157: 196–217. doi:10.1016/s0377-2217(03)00069-9. CiteSeerX: 10.1.1.62.8919.

[5] Erdem, E. S. (2014). Ses sinyallerinde duygu tanıma ve geri erişimi (Master's thesis, Başkent Üniversitesi Fen Bilimleri Enstitüsü).

[6] Topal, C. (2017). Alan Turing'in Toplumbilimsel Düşünü: Toplumsal Bir Düş Olarak Yapay Zekâ. DTCF Dergisi, 57(2).

[7] Cox, D. R., & Oakes, D. (1984). Analysis of survival data. Chapman&Hall, London.

[8] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

[9] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep, 50, 163-170.

[10] Cox, D. R. (1992). Regression models and life-tables. In Breakthroughs in statistics (pp. 527-541). Springer, New York, NY.

[11] Fleming, T. R., & Lin, D. Y. (2000). Survival analysis in clinical trials: past developments and future directions. Biometrics, 56(4), 971-983.

[12] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[13] Huigevoort, C., & Dijkman, R. (2015). Customer churn prediction for an insurance company.

[14] Sirmacek, B. (2007). FPGA ile mobil robot için öğrenme algoritması modellenmesi (Doctoral dissertation).

[15] Altunisik, R. (2015). Büyük Veri: Fırsatlar Kaynağı mı Yoksa Yeni Sorunlar Yumağı mı?. Yildiz Social Science Review, 1(1).

[16] Gor, İ. (2014). Vektör nicemleme için geometrik bir öğrenme algoritmasının tasarımı ve uygulaması (Master's thesis, Adnan Menderes Üniversitesi).

[17] Turkmenoglu, C. (2016). Türkçe Metinlerde Duygu Analizi (Doctoral dissertation, Fen Bilimleri Enstitüsü).

[18] Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: a contingent perspective. Journal of service research, 19(2), 123-141.

[19] Kocadayi, Y., Erkaymaz, O., & Uzun, R. (2017). Yapay Sinir Ağları ile Tr81 Bölgesi Yıllık Elektrik Enerjisi Tüketiminin Tahmini. BİLDİRİ ÖZETLERİ KİTABI, 239.

[20] Albayrak, A. S., & Yilmaz, Ö. G. Ş. K. (2009). Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1).

[21] Guneren, H. (2015). Destek vektör makineleri kullanarak gömülü sistem üzerinde yüz tanıma uygulaması (Doctoral dissertation).

[22] Ozkan, H. (2013). K-Means Kümeleme ve K-NN Sınıflandırma Algoritmalarının Öğrenci Notları Ve Hastalık Verilerine Uygulanması.

[23] Brownlee J. (2016). A Gentle Introduction to XGBoost for Applied Machine Learning. Recieved from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ Accessed 04 January 2019

[24] Collett, D. (1994). Modelling survival data. In Modelling Survival Data in Medical Research (pp. 53-106). Springer US.

[25] Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. Journal of Interactive Marketing, 20(3-4), 16-29.

[26] Wong, K. K. K. (2011). Using cox regression to model customer time to churn in the wireless telecommunications industry. Journal of Targeting, Measurement and Analysis for Marketing, 19(1), 37-43.

[27] Martinsson, E. G. I. L. (2016). Wtte-rnn: Weibull time to event recurrent neural network (Doctoral dissertation, Master's thesis, University of Gothenburg, Sweden).

[28] Castanedo, F., Valverde, G., Zaratiegui, J., & Vazquez, A. (2014). Using deep learning to predict customer churn in a mobile telecommunication network.

[29] Wangperawong, A., Brun, C., Laudy, O., & Pavasuthipaisit, R. (2016). Churn analysis using deep convolutional neural networks and autoencoders. arXiv preprint arXiv:1604.05377.

[30] Spanoudes, P., & Nguyen, T. (2017). Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors. arXiv preprint arXiv:1703.03869.