Karaelmas Fen ve Mühendislik Dergisi
Journal home page: http://fbd.beun.edu.tr

Review
Received / Geliş tarihi : 30.01.2017
Accepted / Kabul tarihi : 08.03.2017

Review of Customer Churn Analysis Studies in Telecommunications Industry

Telekomünikasyon Sektöründe Müşteri Ayrılma Tahmin Analizi Çalışmaları Derlemesi

Fatih Kayaalp

Düzce Üniversitesi, Bilgisayar Mühendisliği Bölümü, Düzce, Turkey

Abstract

Churn Analysis is one of the world wide used analysis on Subscription Oriented Industries to analyze customer behaviors to predict the customers which are about to leave the service aggrement from a company. It is based on Data Mining methods and algorithms and become so important for companies in today's commercial conditions as gaining a new customer's cost is more than retaining the existing ones.

The paper reviews the releveant studies on Customer Churn Analysis on Telecommunication Industry in literature to present a general information to readers about the frequently used data mining methods used, results and performance of the methods and shedding a light to further studies. To keep the review up to date, studies published in last five years and mainly last two years have been included.

Keywords: Churn analysis, Data mining, Telecommunications

Öz

Müşteri Ayrılma Tahmin Analizi, dünya çapındaki müşteri odaklı sektörlerdeki şirketlerin müşterilerinin davranışlarını analiz ederek, bu müşterilerden hizmet almayı bırakmayı düşünenleri tahmin etmeye yönelik olarak kullandıkları bir inceleme şeklidir. Veri madenciliği temelli bu analiiz yöntemi, günümüzdeki ticari şartlarda yeni müşteri kazanmanın eldekini tutmaya göre daha maliyetli olması dolayısıyla çok daha önemli bir hale gelmiştir.

Sunulan çalışmada, literatürde bulunan haberleşme sektörüne yönelik yapılmış Müşteri Ayrılma Tahmin Analizi çalışmaları, bu çalışmalarda sıklıkla kullanılan veri madenciliği yöntemleri, elde edilen sonuçlar ve performansları hakkında bilgi vermek ve ileriye yönelik çalışmalara ışık tutmak amaçlanmıştır. Derlemenin güncel olması için de son beş yıldaki yayınlar ve ağırlıklı olarak da son iki yıldaki çalışmalara yer verilmiştir.

Anahtar Kelimeler: Müşteri ayrılma tahmin analizi, Veri madenciliği, Haberleşme

1. Introduction

Studies revealed that gaining new customers is 5 to 10 times costlier than keeping existing customers happy and loyal in today's competitive conditions, and that an average company loses 10 to 30 percent of customers annually (Kotler 2009). Many companies, being aware of this fact, are engaged in satisfying and retaining the customers.

Especially in the subscription oriented industries, such as telecommunications, banking, insurance, and in the fields of customer relationship management, etc., companies working with numerous customers, the revenues of the companies are provided by the payments made by these customers periodically. It is very important to be able to keep customers satisfied in order to be able to sustain this revenue with the least expenditure cost.

The objectives of this study are:

- Reviewing the relevant studies about churn analysis on telecommunications industry presented in the last five years, particularly in the last two years, and introducing these up-to-date studies in the literature,
- Determining the data mining methods frequently used in churn implementations,
- Shedding a light on methods that can be used in further studies.

2. Data Mining and Customer Churn Analysis

In today's technological conditions, new data are being produced by different sources in many sectors. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out these hidden information, various analyses should be performed using data mining, which consists of numerous methods.

The Churn Analysis aims to predict customers who are going to stop using a product or service among the customers. And, the customer churn analysis is a data mining based work that will extract these possibilities. Today's competitive conditions led to numerous companies selling the same product at quite a similar service and product quality. In the midst of this competition, the cost of gaining new customers is more than retaining the existing customers. For this reason, existing customers are very valuable.

With the Churn Analysis, it is possible to precisely predict the customers who are going to stop using services or products by assigning a probability to each customer. This analysis can be performed according to customer segments and amount of loss (monetary equivalent). Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be created by calculating the churn rate or customer attrition. In this way, profitability can be increased significantly or the possible damage due to customer loss can be reduced at the same rate (Argüden 2008).

For example, if a service provider which has a total of 2 million subscribers, gains 750.000 new subscribers and losts 275.000 customers; churn rate is calculated as 10%. The customer churn rate has a significant affect on the financial market value of the company. So most of the companies keep an eye on the value of the customer at monthly or quarterly periods (Seker 2016).

Churn can be called as voluntary and involuntary. Voluntary churn occurs when an existing customer leaves the service provider and joins another service provider; but in involuntary churn, customer is asked by the service provider to leave due to reasons like non-payments etc. (Mahajan 2015). Voluntary churn can be sub-divided into: incidental churn and deliberate churn (Gotovac 2010). Incidental churn occurs because of the unplanned changes in the customers' lives like a change in financial conditions, change in living location. Deliberate churn occurs for reasons of technology (customers that want a newer or better technology, price sensitivity, service quality factors, social or physological factors and convenience reasons) (Mattison 2005)

3. Studies of Churn Analysis in Telecommunication Industry

In a study by Gursoy, customers who tend to leave a large company operating in the telecommunication sector in Turkey have been identified to develop special marketing strategies for these customers. Logistic Regression Analysis and Decision Tree classification techniques have been used on a 4-month data set consisting 1000 records with 24 variables, and the results have been presented (Gursoy 2010).

Verbeke and friends have set up their own models on 11 different data sets with 20 variables taken from real customers, and presented the details of the maximum-profit-centered model. In addition, the results obtained by applying commonly used data mining methods (decision trees, neural networks, support vector machines, classification, etc.) have been compared (Verbeke 2012).

In a study by Yabas, a method analyzing the customer churn by using the latest data mining methods to predict the customer churn is described. Various classification methods have been applied on the data set of 100.000 records with 230 variables provided by Orange Telecommunication Company, and the performance results have been summarized (Yabas 2012).

In a study by Kamalraj and Malathi, the performances of the J48 decision tree and C5.0 classification technique have been compared on customer churn analysis on a data set including 3333 customer records with 10 variables. The C5.0 algorithm has been reported to provide better predictions and use lower run-time memory compared to that of J48 algorithm (Kamalraj 2013).

In the churn analysis study by Brandusoiu and Toderean, 4 different core functions have been used in the Support Vector Machines model and performances have been compared by using a data set consisting of 3333 customer records with 21 variables provided by a telecommunications company. And among these models, the one with the polynomial core function has been reported to have the best result by 88.56% (Brandusoi 2013).

In Olle and Cai's study, a customer churn analysis has been performed in WEKA environment with a hybrid model using a 6-month data set, having 2000 subscriber records

with 23 variables, given by an Asian telecommunication company. Logistic Regression (LR) for classification, and Voted Perceptron (VP) for estimation has been used in the model. According to the results obtained, the predictive power of the hybrid model has been reported to be better than of individual methods (Olle 2014).

Forhad and friends have performed a churn analysis and presented the results in their study using rule-based classification method on a 26-month data set, having 6938 records, and 880 phone numbers (Forhad 2014).

In the study by Amin and friends, the application of the coarse set theory on single-class or multi-class classifiers has been explained, and the results have been compared by applying detailed, genetic, overlay and LEM2 algorithms on single- and multi-class models on an sample data set with 12 variables, obtained from a public source. The genetic algorithm implementation has been reported to provide the best result (Amin 2014).

The study presented by Gok et al, proposes a two-phase solution utilizing data mining techniques which are time series clustering and classification techniques. The first phase runs for each behavior set (time series) and the second phase is classification algorithms process on enriched and interpreted raw data to predict whether the customer will churn or not. Support vector machines (SVM) and Recursive PARTitioning (RPART) are used in performance comparison. The performance of the whole two phased algorithm is measured ath the end of the algorithm by taking means of of the performance index at each trial of the second phase. The model has been applied on a data set of 70000 records of 6000 customers with 13 variables from leading internet service provider company Turksat Satellite Communications and Cable TV and the results have been reported (Gok 2015).

The study conducted by Kaur and Mahajan presents a churn analysis performed using R program and J48 decision tree method on an example data set with 21 variables, obtained from "http://www.dataminingconsultant.com/data/churn.txt" and "http://www.sgi.com/tech/mlc- /db/" resources, in the telecommunications sector. Various functions have been tested on the customer attrition factor, and churn analysis have been performed using various alternative graph plots offered by the R program (Kaur 2015).

In their study, Hudaib et al. have developed 3 hybrid models to predict customer churn in telecommunications companies, and investigated the performances of these models on a 3-month data set of 500 records with 11 variables, provided by Jordan telecom company. In the first model, k-means algorithm for data filtering and Multilayer Perceptron Artificial Neural Networks (MLP-ANN) method for estimation has been used, the MLP-ANN and hierarchical clustering method has been used in the second model, and the MLP-ANN and Self Organizing Map (SOM) method has been used in the third model. They have compared the results of hybrid models with those of C4.5 and MLP-ANN methods in terms of model accuracy and customer attrition rates, and have stated that all the hybrid models had better results compared to individual models, and the k-means + MLP-ANN model had the best result among hybrid models (Hudaib 2015).

Yildiz has conducted a study to predict the customer churn using data mining classification techniques. In order to reduce the run-time of the classification techniques and to increase the performance, they have reduced the number of features, used different classification techniques and measured their performances. In addition, outlier analysis has been performed to observe the effects on the classification results. These classifications have been tested on 2 different data sets containing 5000 subscribers with 20 variables and 51306 subscribers with 172 variables, and Recall Ratio and Precision Ratio have been used as the performance criteria (Yildiz 2015).

Backiel and friends discusses a homophily-based customer churn analysis implementation, described as the tendency of individuals to demonstrate similar behavior to those in their social network environment, as well as investigating the details of phone call records of individuals. They have used a data set of 6 months' call records of 1 million customers with 111-variable including data such as which customer had a conversation with whom, how many times, and for how long, which is used to depict their social network provided by a GSM operator. They have stated that the test results of these combined factors were more successful than of the individual test cases (Backiel 2015).

In the study by Dahiya and Bhatia, two different models based on decision trees and logistic regression were constructed to perform customer churn analysis on 3 types of data sets including 50 records with 10 variables, 200 records with 50 variables, and 608 records with 100 variables, in WEKA environment. As a result of the comparisons made in accordance with the results obtained, the decision tree-based model has been reported to provide better estimation results (Dahiya 2015).

In the study by Dalvi, data mining and machine learning techniques have been utilized, and logistic regression and decision tree-based customer churn prediction models have been presented using R program on a customer data set taken from telecommunications company. They have used 19 features obtained from the customers' call records, compared their results, and indicated that the decision-trees method had a better predictive accuracy (Dalvi 2016).

In their study, Gordini and Veglio have focused on the churn analysis in the determination of marketing strategies. Although it is not directly related to the field of telecommunications, this study can potentially be adapted to this field. A Support Vector Machine based solution model (SVMauc) based on the AUC parameter selection technique has been designed in the study, which has been carried out on a data set of 80000 customers' records with 24 variables, obtained from an Italian company that sells various products over the Internet. The performance of this model has been compared with logistic regression, neural networks and classical support vector machines, and results have been reported (Gordini 2016).

In their study, Yihui and Chiyu have stated that they have developed a feature extraction method, called FE_RF&T, and a variable/feature selection method, called OOPM. They found that the OOPM method, which has been used for feature selection, has advantageous compared to the Random Forest method, and that the FR_RF&T method has been more successful compared to the PCA method, according to the results of applying the proposed model on a sample data set including 16920 records with 22 variables, provided by the China Mobile Corp. (Yihui 2016).

In a study by Brandușoiu et al., confusion matrix values, gain measure and ROC curve results were investigated by developing models based on neural networks, support vector machines and Bayesian network methods on a data set of 3333 customers with 21 variables, received from California University. It has been observed that Bayesian Network, Multi-Layer Perceptron, and SVM have 99.10%, 99.55%, and 99.70% accurate predictions respectively (Brandusoiu 2016).

The study by Oskarsdottir et al. describes the implementation of customer churn analysis through a learning and classification-based model on the records of communication in social networks, indicating the close circle of the customers that they have had frequent conversations with. Information on implementation of relational learning and relational classifiers on 7 different data sets with an average of 1 million records as well as the comparison results have been given in the study (Oskarsdottir 2016).

A study by Yu et al. proposes an algorithm based on back-propagation neural networks using piecewise classification optimization, called PBCCP. The results of tests performed using BP, PSO-BP, and PBCCP on actual data, inluding hundreds of records with 7 variables provided by China Mobile, have been compared. And the churn analyses performed using the PBCCP algorithm and BP neural networks with optimized weight and threshold values have been shown to yield better results (Yu 2016).

AlOmari and Hassan describes a study using Rules 6-C algorithm which has never been used in a customer churn analysis before. After segmentation and feature selection process, the performance measurement obtained using the 6th algorithm of the RULES Family (RULES 6-C) has been anlyzed (AlOmari 2016).

Amin and his friends proposes an intelligent rule-based decision-making technique, based on rough set theory (RST), to extract important decision rules related to customer churn and non-churn. Many experiments are carried out to evaluate the performance of the proposed RST based CPP (Customer Churn Prediction) approach by using four rule-generation mechanisms, namely, the Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 algorithm. The results show that RST-GA is the most efficient technique for extracting implicit knowledge in the form of decision rules from the publicly available, benchmark telecom dataset which contains 3333 records with 11 variables (Amin 2016).

In her PhD thesis, Tanneedi presents a study with BigData analytics and machine learning to identify churn. She tries to predict customer churn using Big Data analytics, namely a J48 decision tree on a Java based benchmark tool, WEKA. Three different datasets from various sources were examined; first one includes Telecom operator's six months aggregate active and churned users' data usage volumes; second one includes globally surveyed data and third dataset includes of individual weekly data usage analysis of 22 android customers along with their average quality, annoyance and churn scores by accompanying theses. Statistical analyses and J48 Decision trees were applied on these datasets. From the statistics of normalized volumes, autocorrelations were small owing to reliable confidence intervals, but confidence intervals were overlapping and close by, therefore no much

significance could be noticed, henceforth no strong trends could be observed. From decision tree analytics, decision trees with 52%, 70% and 95% accuracies were achieved for three different data sources respectively (Tanneedi 2016).

Mahajan and Som present a study on analyzing customer behaviors on the customers' pre-paid recharge data, voice and SMS usage data to identify patterns in user behavior for intelligent and targeted promotions and churn prediction over the dataset taken from BSNL telecommunications company in India. The number of records of the dataset is not clear as data about different types are included. But generally 25 variables on customer details, recharge details, outgoing and incoming voice calls and sms sent are used. And a logistic model on predicting customer churn has been offered (Mahajan 2016).

A study by Li and his friends list the difficulties for customer churn prediction and then summarizes the method they have proposed as a new supervised one-side sampling technique to pre-process the imbalanced data set which was taken from a Chinese telecom company having 2,7 million records with big number of variables. After applying K-means method to cluster the data set, one-sided sampling is applied in each cluster to remove noise and redundant negative samples. Random forest method is used for dimensional reduction and selecting the required variables. By this way, 9 daily variables are used in the analysis. C5.0 decision tree is used as the classifier to predict customer churn. It is reported that a precision ratio of 80,42% with a recall ratio of 52,43% has been achieved (Li 2016).

Esteves reports a study about churn analysis by using KNN, Naïve Bayes, C4.5, Random Forest, AdaBoost and ANN over a dataset provided by WeDo telecom company of 100 thousand calls from 160 clients between 30 June 2012 and 31 January 2013 of 14 variables. Multiple approaches to predict customer churn are compared. The models have been validated by using 10-fold cross validation with 3 repeats. A hybrid sampling method (SMOTE) has been used to balance the data set. The random forest model with the highest ROC value of 0.9915 and Sensitivity value of 0.9110 has performed the best among all others (Esteves 2016).

Petkovski and his friends gives information about phases of churn analysis which are understanding the business; selection, analysis and data processing; implementing various algorithms for classification; evaluation of the classifiers and choosing the best one for prediction. The results of the churn analysis performed on a data set provided by a telecom company from Macedonia by using decision trees, KNN, logistic regression, and naive Bayes are given. The data set covers 28 months period (approximately 34 million records) with 68 variable that are categorized as demographic, contract and customer behavior attributes of 22461 customers. The highest accuracy has been performed by logistic regression as 94,351% (Petkovski 2016).

Umayaparvathi reports a study that explores the application of data mining techniques in predicting the likely churners and attribute selection on identifying the churn. It also compares the efficiency of several classifiers (like Logistic regression, KNN, Random Forest, SVM, Ridge classifier, Decision Tree, Gradient Boosting) and lists their performances on two real telecom datasets on performance metrics as confusion matrix, accuracy, precision and recall, F1-score. It is also reported that the Gradient Boost classifier has outperformed others (Umayaparvathi 2016).

Stripling presents a study on churn called ProfLogit that explicitly takes profit maximization concerns into account during the training step, rather than the evaluation step. The technique is based on a logistic regression model which is trained using a genetic algorithm (GA). By means of an empirical benchmark study applied to real-life data sets, it is showed that ProfLogit generates substantial profit improvements compared to the classic logistic model for many data sets. In addition, profit-maximized coefficient estimates differ considerably in magnitude from the maximum likelihood estimates (Stripling 2015).

Hossain's study on churn, aims to evaluate various SVM kernels to predict churners on an imbalanced distribution of churners and non-churners data set. The experiment was carried out on a telecommunication data taken by random sampling. The data set includes 12000 records with a total variable number of 57 on customer call, SMS/MMS, EDGE, VAS, revenue and recharge related information. The study shows that linear kernel outperforms commonly used Radial Basis Function(RBF), sigmoid and polynomial kernels (Hossain 2015).

In a study by Rodan, use of an Echo State Network (ESN) with a Support Vector Machine (SVM) training algorithm for predicting customer churn in telecommunication companies is presented. The proposed approach has been trained and tested on two datasets which has 3333 customers' records with 16 variables and 5000 customers' records with 11 variables. The comparison of the proposed

model with other popular machine learning models like Classical SVM with RBF kernel, Multilayer Perceptron (MLP) Neural Network with backpropagation learning algorithm, k-Nearest Neighbour (IBK), Naive Bayes (NB) and C4.5 Decision Trees algorithm has been performed. And the experiment results show that ESN with SVM readout outperforms all other machine learning models used (Rodan 2015).

Mohanty proposes a churn study employing Counter Propagation Neural Networks (CPNN), Classification and Regression Trees (CART), J48 and fuzzyARTMAP to predict customer churn and non-churn in telecommunication sector on a data set taken from Indian Telecommunication Service Industry having 125 records and 5 variables (Mohanty 2015).

Li and friends proposes 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 (5000 records with 20 variables) and Orange (50000 records with 230 variables) public data sets with ROC curves and AUC value as the evaluation index of experiments, comparison of the experimental results show that the presented method for telecom customer churn prediction has the stable promotion effect (Li 2014).

Qureshi reports a study on churn prediction on historical data by using some of the well-known algorithms which are Regression analysis, Decision Trees, KNN and Artificial Neural Networks (ANNs) on a data set containing 106000 customers traffic data (outgoing, incoming, voice, SMS (Short Message Service), data) of 3 months obtained from Customer DNA website. The use of re-sampling method in order to solve the problem of class imbalance has also been discussed. The results show that in case of the data set used, decision trees is the most accurate classifier algorithm on identifying potential churners (Qureshi 2013).

Lu and friends presents a study on customer churn prediction and proposes the use of boosting to enhance a customer churn prediction model. Unlike most research that uses boosting as a method to boost the accuracy of a given basis learner, this paper tries to separate customers into two clusters based on the weight assigned by the boosting algorithm. As a result, a higher risk customer cluster has been identified. Logistic regression has been used in this research as a basis learner, and a churn prediction model has been built on each cluster, respectively. A data set including 7181 records of 698 customers with 21 variables has been used in the experiments. The results have been compared with a single logistic regression model. Experimental evaluation reveals that boosting also provides a good separation of churn data. Thus, boosting has been suggested for churn prediction analysis (Lu 2014).

Idris proposes a churn prediction approach that exploits the discriminative feature selection capabilities of minimum redundancy and maximum relevance in the first step, leading to enhanced feature-label association and reduced feature set. The diverse ensemble of different base classifiers is then applied as a predictor in a second step. Final predictions are computed based on majority voting Random Forest, Rotation Forest and KNN, that ultimately leads to predicting churners from telecom datasets with higher accuracy on two data sets that have 50000 records with selected 39 variables and 40000 records with selected 20 variables, respectively. Simulation results are evaluated using sensitivity, specificity, area under the curve (AUC) and Q-statistic based measures. The results show that the proposed approach efficiently models the challenging problem of telecom churn prediction, by effectively handling the large dimensionality and extending useful features to a diverse, majority voting based ensemble (Idris 2014).

Coussement and friends present a study showing the affect of data preparation treatment (DPT) on the performance of churn prediction, leading to improvements of up to 14.5% measured by AUC and 34% in TDL through experiments by using eight state-of-the-art data mining techniques which are Bagged Cart, Bayesian Network, J4.8 decision tree, Multilayer perceptron neural network, Naive Bayes, Random Forest, Radial basis kernel support vector machine and Stochastic gradient boosting. A data set of 30104 customers' records with a variable number varying from 49 to 92 according to the feature selection method used has been used in the tests. As a result, the importance of the DPT is highlighted (Coussement 2016).

4. Conclusions

Data mining has gained a significant place in the world recently and has an expanding areas of use. Data mining studies are mainly carried out in the fields of education, marketing, banking, stock exchange, medicine, and in the telecommunications sector particularly.

Considering the overall studies reviewed, it is seen that almost all studies have performed training and testing on

Table 1. Classifications of the studies investigated.

Reference Number	Method Used	Data Set Used
Gursoy 2010	Logistic Regression, Decision Trees	1000 records, 24 variables
Verbeke 2012	Decision Trees, Neural Networks, Support Vector Machines, Classification	11 different data sets, having 20 variables
Yabas 2012	Classification	100000 records, 230 variables
Kamalraj 2013	Decision Tree, Classification	3333 customer records, 10 variables
Brandusoiu 2013	Support Vector Machines	3333 customer records, 21 variables
Qureshi 2013	Regression analysis, Decision Trees, KNN and ANNs	106000 customers traffic data records of 3 months, variable number is not clearly identified
Olle 2014	Logistic Regression, Voted Perceptron	2000 customer records, 23 variables
Forhad 2014	Rule-Based Classification	6938 records, 4 variables
Amin 2014	Genetic / Overlay / LEM2 algorithms	3333 records, 12 variables
Li 2014	Cluster stratified sampling logistic regression model	Two datasets having 5000 records with 20 variables and 50000 attributes with 230 variables
Lu 2014	Logistic regression	7181 records of 698 customers with 21 variables
Idris 2014	Random Forest, Rotation Forest and KNN	Two data sets: 50000 records with selected 39 variables and 40000 records with selected 20 variables
Gok 2015	Time Series Clustering and Classification	6000 customers, 70000 records, 13 variables
Kaur 2015	Decision Trees	Unspecified number of records, 21 variables
Hudaib 2015	Neural Networks, Decision Trees, Hybrid Model	5000 customer records, 11 variables
Yıldız 2015	Classification	5000 customer records, 20 variables; 51306 customer records, 172 variables
Backiel 2015	Hybrid Model, using Social Networks	1 million customer records, 111 variables
Dahiya 2015	Logistic Regression, Decision Trees	50 records, 10 variables; 200 records, 50 variables; 608 records, 100 variables
Stripling 2015	Logistic regression, Genetic Algorithm	Six data sets with a record number varying from 889 to 94718 with a variable value of 9 to 37
Hossain 2015	SVM	12000 records with a total number of 57 variables
Rodan 2015	Echo State Network (ESN) with a Support Vector Machine (SVM), Classical SVM with RBF kernel, Multilayer Perceptron (MLP) Neural Network with backpropagation learning algorithm, k-Nearest Neighbour (IBK), Naive Bayes (NB) and C4.5 Decision Trees	Two datasets having 3333 customers' records with 16 variables and 5000 customers' records with 11 variables
Mohanty 2015	Counter Propagation Neural Networks (CPNN), Classification and Regression Trees (CART), J48 and fuzzyARTMAP	Data set having 125 records with 5 variables
Dalvi 2016	Logistic Regression, Decision Trees	Undefined record number, 19 variables

Table 1. Cont.

Reference Number	Method Used	Data Set Used
Gordini 2016	Support Vector Machines, Logistic Regression, Neural Networks	80000 customer records, 24 variables
Yihui 2016	Hybrid Model	16920 records, 22 variables
Branduşoiu 2016	Neural Networks, Support Vector Machines, Bayesian Networks	3333 customer records, 21 variables
Oskarsdottir 2016	Hybrid Model, using Social Networks	7 different data sets, having 1 million records approximately
Yu 2016	Hybrid Model	Unspecified number of records, 7 variables
AlOmari 2016	Rules Family Algorithm 6 (RULES 6-C)	10000 customer records, 8 variables
Amin 2016	Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA), LEM2 algorithm	3333 records, 11 variables
Tanneedi 2016	Big Data analytics (statistical analysis), J48 decision tree	4106 records, 16 variables, 8 weeks' data usage of 22 Android users, 788 customers survey on 13 variables
Mahajan 2016	Logistic model	Unspecified number of records, 25 variables
Li 2016	Decision Tree, Random Forest	Data set of 2.7 million records with 9 selected variables
Esteves 2016	KNN, Naïve Bayes, C4.5, Random Forest, AdaBoost and ANN	100 thousand calls from 160 clients with 14 variables
Petkovski 2016	Decision trees, KNN, logistic regression, and naive Bayes	Approximately 34 million records of 22461 customers with 68 variables
Umayaparvathi 2016	Logistic regression, KNN, Random Forest, SVM, Ridge classifier, Decision Tree, Gradient Boosting	Dataset 1: 70831 records with 74 variables Dataset 2: 333 records with 18 variables
Coussement 2016	Bagged Cart, Bayesian Network, J4.8 decision tree, Multilayer perceptron neural network, Naive Bayes, Random Forest, Radial basis kernel support vector machine and Stochastic gradient boosting	A data set of 30104 customers' records with a variable number varying from 49 to 92

real data provided by different companies operating in the telecommunications industry.

Decision trees, Support Vector Machines, and Hybrid Models, which consist of the use of different methods together, are among the most frequently used ones.

There are many feature selection methods in the literature. It is seen as one of the important things that should be considered carefully as they have significant effect on the performance of the analysis.

The prominent studies in the last two years have investigated the social networks, which is defined as the customers' own close network, regdarding the behavioral influence on customers to the extent of termination of service. And, it has been observed that the addition of users' social networks into analysis improves the performance in predicting the customers churn.

The studies examined show that most of the institutions and organizations are focused on the customer/user behavior analysis today. It will be more realistic for them to make future plans based on these analyses. Moreover, the diversification of the fields on which data mining is used will also provide great benefits to institutions and organizations, as well as to the people who receive services and products from these institutions and organizations in Turkey (Savas 2012).

It is suggested further studies to be conducted by developing hybrid models that are based on communications of individuals with users in their social networks. This will probably give better estimation results.

5. References

AlOmari, D., Hassan, M.M. 2016. Predicting Telecommunication Customer Churn Using Data Mining Techniques. *9th International Conference on Internet and Distributed Computing Systems*, 167-178.

Amin, A., Khan, C., Ali, I., Anwar, A. 2014. Customer Churn Prediction in Telecommunication Industry: with and without counter-Example. *European Network Intelligence Conference,* 134-137.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., Huang, K. 2016. Customer Churn Prediction in Telecommunication Sector using Rough Set Approach. *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2016.12.009, 2016:1-21.

Argüden Y., Erşahin B. 2008. Veri Madenciliği: Veriden Bilgiye, Masraftan Değere. *ARGE Danışmanlık,* ISBN: 978-975-93641-9-9 1. Basım.

Backiel, A., Verbinnen, Y., Baesens, B., Claeskens, G. 2015. Combining Local and Social Network Classifiers to Improve Churn Prediction. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 651-658.

Brandusoiu, I., Toderean, G. 2013. Churn Prediction In The Telecommunications Sector Using Support Vector Machines. *Annals Of The Oradea Un., Fascicle Manag. and Tech. Eng.,* 1: 19-22.

Brandusoiu, I., Toderean, G., Beleiu, H. 2016. Methods for churn prediction in the pre-paid mobile telecommunications industry. *International Conference on Communications (COMM),* 97-100.

Coussement, K., Lessmann, S., Verstraeten, G. 2016. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decis. Supp. Sys.,* 2016, http://dx.doi.org/10.1016/j.dss.2016.11.007.

Dahiya, K., Bhatia, S. 2015. Customer Churn Analysis in Telecom Industry. *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO),* 1-6.

Dalvi, PK., Khandge, SK., Deomore, A. 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *Symposium on Colossal Data Analysis and Networking (CDAN),* DOI: 10.1109/CDAN.2016.7570883, 1-8.

Esteves, G., Mendes-Moreira, J. 2016. Churn Perdiction in the Telecom Business, *The eleventh International Conference on Digital Information Management (ICDIM 2016),* 254-259.

Forhad, N., Hussain, S., Rahman, RM. 2014. Churn Analysis: Predicting Churners. *Ninth International Conference on Digital Information Management (ICDIM),* 237-241.

Gok, M., Ozyer, T., Jida, J. 2015. A Case Study for the Churn Prediction in Turksat Internet Service Subscription. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 1220-1224.

Gordini, N., Veglio, V. 2016. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. I*ndust. Mark. Manag.*, 2016 (1):1-8.

Gotovac, S. 2010. Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services. *Automatika,* 51 (3): 275-283.

Gursoy, UTŞ. 2010. Customer Churn Analysis in Telecommunication Sector. İstanbul *Un. J. School Business Admin.*, 39 (1): 35-49.

Hossain, M.M., Miah, M.S. 2015. Evaluation of Different SVM Kernels for Predicting Customer Churn, *International Conference on Computer and Information Technology (ICCIT),* 1-4.

Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., Faris, H. 2015. Hybrid Data Mining Models for Predicting Customer Churn. *Int. J. Comm. Netw. Sys. Sci.*, 8: 91-96.

Idris, A., Khan, A. 2014. Ensemble based Efficient Churn Prediction Model for Telecom, *12th International Conference on Frontiers of Information Technology*, 238-244.

Kamalraj, N., Malathi, A. 2013. Applying Data Mining Techniques in Telecom Churn Prediction. *Int. J. Adv. Res. Comp. Sci. Softw. Eng.,* 3 (10): 363-370.

Kaur, M., Mahajan, P. 2015. Churn Prediction in Telecom Industry Using R. *Int. J. Eng. Tech. Res. (IJETR),* 3 (5): 46-53.

Kotler, P., Keller, K. L. 2009. Marketing Management. *Pearson Prentice Hall.*

Li, H., Yang, D., Yang, L., Lu, Y., Lin, X. 2016. Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction, *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom),* 163-169.

Li, P., Li, S., Bi, T., Liu, Y. 2014. Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression, *International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things,* 282-287.

Lu, N., Lin, H., Lu, J., Zhang, G. 2014. A Customer Churn Prediction Model in Telecom Industry Using Boosting, *IEEE Trans. On Indust. Inf.,* 10 (2): 1659-1665.

Mahajan, R., Som, S. 2016. Customer Behavior Patterns Analysis in Indian Mobile Telecommunications Industry, *International Conference on Computing for Sustainable Global Development (INDIACom)*, 1165-1169.

Mahajan, V., Misra, R., Mahajan, R. 2015. Review of Data Mining Techniques for Churn Prediction in Telecom. *J. Inf. Org. Sci. (JIOS)*, 39 (2): 183-197.

Mattison, R. 2005. The Telco Churn Management Handbook. *XiT Press,* Oakwood Hills, Illinois.

Mohanty, R., Rani, JK. 2015. Application of Computational Intelligence to Predict Churn and Non-Churn of Customers in Indian Telecommunication, *International Conference on Computational Intelligence and Communication Networks (CICN)*, 598-603.

Olle, GDO., Cai, S.Q. 2014. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *Int J. e-Educ., e-Business, e-Manag. e-Learn.,* 4 (1): 55-62.

Oskarsdottir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., Vanthienen, J. 2016. A Comparative Study of Social Network Classifiersfor Predicting Churn in the Telecommunication Industry. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 1-8.

Petkovski, AJ., Stojkoska, BLR., Trivodaliev, KV., Kalajdziski, SA. 2016. Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia, *24th Telecommunications forum TELFOR*, 1-4.

Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., Rehman, A. 2013. Telecommunication Subscribers' Churn Prediction Model Using Machine Learning, *Eighth International Conference on Digital Information Management (ICDIM),* 131-136.

Rodan, A., Faris, H. 2015. Echo State Network with SVM-readout for Customer Churn Prediction, *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),* 1-5.

Savas, S., Topaloglu, N., Yılmaz, M. 2012. Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri. **İstanbul Ticaret Ün.** *Fen Bil. Derg.,* 11 (21): 1-23.

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

Stripling, E., Broucke, S., Antonio, K., Baesens, B., Snoeck, M. 2015. Profit Maximizing Logistic Regression Modeling for Customer Churn Prediction, *International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10.

Tanneedi, N. N. P. P. 2016. Customer Churn Prediction Using Big Data Analytics. *PhD Thesis from Blekinge Institute of Technology*.

Umayaparvathi, V., Iyakutti, K. 2016. Attribute Selection and Customer Churn Prediction in Telecom Industry, *International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 1-7.

Verbeke, W., Dejaeger, K., Martens, D., Hur. J., Baesens, B. 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European J. Operat. Res.*, 218 (1): 211–229.

Yabas, U., Cankaya, H. C., Ince, T. 2014. Customer Churn Prediction For Telecom Services. *IEEE 36th Annual Computer Software and Applications Conference (COMPSAC),* 358-359.

Yihui, Q., Chiyu, Z. 2016. Research of Indicator System in Customer Churn Prediction for Telecom Industry. *11th International Conference on Computer Science & Education (ICCSE*), 123-130.

Yildiz, M., Albayrak, S. 2015. Telekomünikasyon Sektöründe Müşteri Ayrılma Tahmini. *23th Signal Processing and Communications Applications Conference (SIU),* 256-259.

Yu, R., An, X., Jin, B., Shi, J., Move, O.A., Liu, Y. 2016. Particle classification optimization-based BP network for telecommunication customer churn prediction. *Neural Comput. & Applic.*, DOI: 10.1007/s00521-016-2477-3, 2016:1-14.