SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ

Sakarya University Journal of Science
SAUJS

ISSN 1301-4048 e-ISSN 2147-835X Period Bimonthly Founded 1997 Publisher Sakarya University
http://www.saujs.sakarya.edu.tr/

Title: The Attitudes of the Telecommunication Customers in the COVID-19 Outbreak: The Effect of the Feature Selection Approach in Churn Analysis

Authors: Handan DONAT, Saliha KARADAYI USTA

Recieved: 2022-02-22 00:00:00

Accepted: 2022-04-27 00:00:00

Article Type: Research Article

Volume: 26
Issue: 3
Month: June
Year: 2022
Pages: 530-544

How to cite
Handan DONAT, Saliha KARADAYI USTA; (2022), The Attitudes of the Telecommunication Customers in the COVID-19 Outbreak: The Effect of the Feature Selection Approach in Churn Analysis. Sakarya University Journal of Science, 26(3), 530-544, DOI: 10.16984/saufenbilder.1077229
Access link
http://www.saujs.sakarya.edu.tr/tr/pub/issue/70993/1077229

Sakarya University Journal of Science 26(3), 530-544, 2022

The Attitudes of the Telecommunication Customers in the COVID-19 Outbreak: The Effect of the Feature Selection Approach in Churn Analysis

Handan DONAT[1], Saliha KARADAYI USTA*[1]

Abstract

Today's rising cutting-edge technology requirements and competitive environment in telecommunication industry has gained a remarkable importance due to the COVID-19 pandemics in terms of high need of information sharing and remote communication necessity. Telecommunication companies conduct significant analyses by highlighting that the customer data is the most valuable information. Besides, they obtain results emphasizing that acquiring new customers is costlier than retaining the existing ones. Therefore, the companies are willing to determine the important customer features in order to understand why they shift to the other telecommunication service providers. Hence, this study aims to conduct a churn analysis by feature selection approach with large volumes of telecommunication customer data in order to present what kind of customer behaviors and qualifications exist. Since there is a huge amount of data in this field, data mining is a vital requirement. The performance outputs were observed, and the features carrying these outputs to the highest value were identified. The data collection and analysis were carried out in mid-2019, and the same data collection and analysis were carried out again at the beginning of 2021, and these before and after results were compared. In addition, a comparison was made with the results obtained by the other churn analysis studies. This paper contributes to the practitioners by presenting the most important customer features in telecom customer churn, and a new approach in performance evaluation have been proposed specific to the telecommunication market with the industry experts' guidance as a theoretical contribution.

Keywords: Telecommunication, customer churn, churn analysis, data mining, machine learning

1. INTRODUCTION

Communication is a need for human beings, and it has gained different dimensions with the development of technology day by day [1-3]. Especially with the COVID-19 pandemics, remote access and online communication have been the only option to conduct business, and sharing the same environment was forbidden occasionally. Therefore, digitalization has quickly entered to the individuals' lives, and especially for the companies,

* Corresponding author: salihakaradayiusta@gmail.com
[1] İstinye University, Industrial Engineering Department,
E-mail: handanakterazi@gmail.com
ORCID: https://orcid.org/0000-0002-8348-4033, https://orcid.org/ 0000-0002-8006-0606

it has been inevitable to adapt to the digital environment.

The sector reports of the post-COVID-19 telecom supply chains have announced that the telecom industry has come to the fore with the digital technologies as the most used management tool of goods and services with a rate of 61%. Compared to other sectors, it has been effective in determining the impacts of the pandemics, and according to the customer surveys, 67% of the customers have stated that the telecom sector showed high performance during the pandemics [4]. In order to eliminate the negative effects of the COVID-19 pandemics on the telecom industry, companies have focused in workforce, operations and supply chain, communication strategies, customer data and revenue management [5]. Customer data has been significant as it is of vital importance for the telecom industry [6]. Since the customer information modeling creates an important competitive advantage [7, 8], the new researches discussed the impact of the pandemics in order to observe the changing customers' behaviors in detail.

The level of competition has increased with the technical progress and increasing number of operators in the telecommunications sector. The main strategies that the companies employ to generate more revenue are: acquiring new customers, retaining existing customers, and increasing the time duration and the loyalty of the customers by keeping them in the existing company. The third strategy has been found to be the most profitable strategy. For these reasons, companies should reduce the customer movement from one service provider to another which stands for the "customer churn" [9].

Customer churn is addressed in many industries such as telecommunication [10-13], peer-to-peer lending market [14], retail industry [15, 16], and banking [17, 18]. As it is obvious, the telecom industry mostly conducts churn analysis in order to keep the customers in the existing company by analyzing the customer specifications and behaviors.

The telecom churn analysis techniques are decision trees [19-21], logistics regression [22, 23], Bayesian networks [24], artificial neural networks [25-28], support vector machine [29, 30], K-means clustering [31, 32], machine learning [10, 12, 13, 18]. Here, the literature review points out to the artificial intelligence algorithms that are needed in processing the big data. Moreover, the telecommunication industry provides services like local and long distance calls, voice, fax, e-mail and other data traffic services. Due to globalization and market conditions, customer relationship management is becoming more and more important. Hence, due to the big amount of customer data, data mining is a requirement to understand business needs, to define the telecommunication model, to use resources efficiently and to increase service quality [33].

The purpose of this paper is implementing a churn analysis by feature selection approach with large amount of customer data in order to illustrate changing direction of customer behaviors. Owing to the existing big data, the machine learning techniques were applied. The performance outputs were recorded, and the customer features bringing these outputs to the highest value were determined. The data collection and analysis were carried out separately in both mid-2019 and at the beginning of 2021, and these two before and after COVID-19 results were compared. Additionally, a comparison was made with the results obtained by the other churn analysis studies as a discussion. Following sections include the methodology part with literature review, classification algorithms, feature selection, performance measurements, application, findings, discussion and conclusion parts.

2. METHODOLOGY

2.1. Literature Review

Data and techniques used in the telecom customer churn analysis vary in the literature. In case of analyzing the payment information, the monthly invoice amount [20, 34, 35] has been the mostly focused factor, while there were a few studies analyzing the billing trend [24]. In addition, call frequency [34], additional service payment information [20], and SMS cost [34] were also taken into consideration. On the other hand, the studies analyzing usage information has been focused on the use of minute [30, 35, 36], usage time intervals [37], usage frequency [38], SMS usage [36], evening and night usage frequency, vocal message

usage, the amount paid for services that require a monthly subscription, internet usage, the amount paid for the internet [38], and the time spent on the internet for international calls [36].

Metrics applied in the studies looking at the line information have been subscription age of the customer [38], payment type (with / without contract) [20, 39], line shifting information, deactivation date, activation date, line opening date, line closing date [37], whether it is 3G or 4G [35], the package used [24].

In the demographic analysis, age and gender [21, 40, 41], income [34], child info and customer profession [21], place of residence (rural / city) [42], education level [39], marital status [40] have been investigated.

In the analysis of customer internal information; billing institution, invoice complaints, weekly number of calls (call center), national and international invoicing [34], customer contact information with the operator [37], customer general complaint information [40] have been addressed. Additionally, customer present value data, customer potential value data, trust, perception, satisfaction, expectation [21], and the device used [30, 39] have been also examined.

2.2. Brief Information about Classification Algorithms

2.2.1. Naive Bayes Classification

In case of deciding whether a given x (x = [x (1), x (2), …, x (L)]$^T \in R^L$) belongs to class S_i or not, and if the independence proposition of Bayes decision theorem is used, then this type of classification is called NB classification. According to the NB decision theory, x belongs to S_i if $P(S_i)\prod_{k=1}^{L} P(x_k|S_i) > P(S_j)\prod_{k=1}^{L} P(x_k|S_j)$ where $P(S_i)$ and $P(S_j)$ are prior possibilities of classes i and j. Thus, the values can be easily calculated from the data set [43]. NB classification conducts good analysis in terms of both accuracy and computation time as a simple and quick statistical prediction algorithm. The NB classifier may be trained to categorize patterns with thousands of attributes and then used to classify thousands of patterns. As a result, NB is the algorithm of choice for data mining and other large classification challenges [44]. The existence (or lack) of a given property of a class (client mix) is assumed to be unrelated to the presence (or absence) of any other property by an NB classifier. The NB classifier outperformed other widely used algorithms in the prediction challenge for the wireless telecommunications industry, and it also increased prediction rates [45].

2.2.2. Logistics Regression

For discriminative probabilistic classification, logistic regression is a frequently used statistical modeling technique [46]. When the dependent variable is binary, this is the predictive regression strategy to use. It's used to describe data and explain the link between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables [47]. It estimates the probability of a particular event that occurs, and the probability calculation as $prob\,(y=1) = \frac{e^{\beta_0+\sum_{k=1}^{k}\beta_k x_k}}{1-e^{\beta_0+\sum_{k=1}^{k}\beta_k x_k}}$. Here y is the dependent variable that indicates whether the event occurred or not (when the event happens y = 1, otherwise y = 0). $x_1, x_2, \ldots, x_k$ are independent inputs. $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients that can be computed using the maximum likelihood technique using the training data [48].

2.2.3. Support Vector Machines

Support Vector Machines (SVM) is a highly preferred algorithm for machine learning by producing significant accuracy with less computation power for the classification duties. The SVM algorithm uses a linear combination of subsets (support vectors) of a training set to discover a hyperplane in an N-dimensional space (N is the number of features) [49]. If input vectors can be non-linearly separated, the support vector machines first map data into a high-resolution (possibly infinite) size feature area that uses a kernel number, and next classify the data by maximum margin. Equation used is $f(\vec{x}) = sgn(\sum_i^M y_i x_i \emptyset(\vec{x}_i, \vec{x}) + \delta)$. While the number of samples in the set of training is M, $\vec{x}_i$ is the vector with support $\vec{x}_i > 0$, ϕ is the kernel function, $\vec{x}_i$ is a sample feature vector that is unknown, and δ is threshold. Parameters $\vec{x}_i$ can be obtained by applying linear constraints to a

convex quadratic programming problem. Kernel functions are frequently used in practice with the polynomial kernel and Gaussian radial elementary functions. Consider the Karush - Kuhn - Tucker case for further information, and selecting any i with $\vec{x}_i > 0$ (supports vectors). In practice, it is, however, safer to use the average value of all support vectors [48].

2.2.4. Artificial Neural Networks

Artificial Neural Networks (ANN) are one of the information processing technology's artificial intelligence applications. It imitates the structure of the human brain and analyzes the existing data and generates new information from the data [50]. The advantages of ANN are being non-linear, ability to design with input and output mappings, possibility of adapt, being tolerant to fault [51]. The input, hidden, and output layers of a feed forward neural network are usually present. A sigmoid function is used to activate artificial neural networks. If an ANN has a hidden layer, the network outputs receive the hidden unit's activation functions by transforming it into a second layer of process components as an $output_{net}(j) = f\left(\sum_{l=1}^{L} w_{jl} f(\sum_{i}^{D} w_{li} x_i)\right), j_1, \dots \dots J.$ f is an activation function, while *D*, *l*, and *j* are the total number of units in the input, hidden, and output layers, respectively. To train ANNs, back propagation or fast back propagation learning techniques are used. [48].

2.3. Feature Selection

While establishing the customer churn model, it is checked how much the variables affect the algorithm on each customer base in order to create the best model. After running the algorithm, one should select the variables that have a greater impact and that provide higher accuracy results. In this study, the variables that affect the algorithm most with the information gain, gain ratio, Gini coefficient and correlation were sorted and the algorithms were run one by one in an order. Moreover, there is a clear advantage of feature selection by reducing the data size and the computational burden, being an easier analysis for the data with smaller data size, by decreasing data complexity, and by increasing class / numerical performance [52, 53]. In the following part, the feature selection metrics will be explained in brief.

2.3.1. Information Gain

Information gain is a method based on entropy having values between 0 and 1 by showing the disorder or uncertainty of the system. If the entropy value approaches 1, it indicates that the system is more regular and specific. Entropy value is calculated by $E = -\sum_{i=1}^{n} \left(log_2 \frac{ns(i)}{N}\right) * \frac{\mathrm{ns}(i)}{N}$ formula where E is entropy, N is the total sample size, n is number of classes, and *ns (i)* is the sample for the i^{th} class [54].

2.3.2. Gain Ratio

At each node of the decision tree, the gain ratio metric is used to select the test feature. This is a variation of information gain that lowers bias by considering the number and size of branches when selecting a feature and adjusting the information gain with the intrinsic information of a split [55]. It favours characteristics with a high number of possible values. C4.5 created the fundamental decision tree induction technique ID3, which employs a known extension of information gain to address this bias [56]. The gain ratio formula is based on the “gain = new ratio – old ratio” equation [57].

2.3.3. Gini Index

The Gini index is a strategy for separating pollution that works with binary, continuous numerical values. Breiman first introduced it in 1984 [58] and it is widely used in algorithms. As in the information gain and gain ratio, gain is calculated for each individual variables for ranking The Gini index is a scale that ranges from 0 to 1, with 0 denoting that all elements belong to one class or that there is only one class, and 1 denoting that the elements are randomly distributed across all classes. A Gini value of 0.5 indicates that items in some classes are evenly distributed. $Gini = 1 - \sum_{i=1}^{n} (p_i)^2$ is the Gini index formula where p_i is the likelihood of an object being categorized into a specific class. When constructing the decision tree, the root node should be the feature with the lowest Gini index. [59].

2.3.4. Correlation

The state of a linear relationship between two random variables is depicted by correlation in statistics. The correlation coefficient is used to determine whether or not there is a linear relationship and, if so, how strong it is. It ranges from -1 to +1 [60]. 1 indicates that for every positive increase in one variable, a fixed proportion increases in the other, whereas -1 indicates that for every positive increase in one variable, a fixed proportion decreases in the other. There is no positive or negative rise for every increase of zero. There is no connection between the two. Furthermore, the correlation coefficient's absolute value indicates the strength of the association. The stronger the association, the higher the number. $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$ is the correlation formula [61].

2.4. Performance Measurements

Following the classification, success is evaluated according to the complexity matrix stated in Table 1.

Here, TP stands for true positive number, correctly defined positive samples, FP stands for false positives (incorrectly recognized negative samples), FN for false negatives (incorrectly identified positive samples), and TN for true negatives (incorrectly identified positive samples), correctly identified negative samples [62].

Table 1
Class Confusion Matrix

		Predicted		Total
		Churn Customer	*Non Churn Customer*	
Actual	*Churn Customer*	TP - True Positive	FN - False Negative	Actual Positive Number
	Non Churn Customer	FP - False Positive	TN - True Negative	Actual Negative Number
Total		Predicted Positive Number	Predicted Negative Number	Total Customer Number

There are many defined performance measurement metrics in the literature. Specificity is the ratio of correctly defined mixed negatives and defined as Specificity = (TN) / (TN+FP). Accuracy measures the overall sampling rate of the number of correctly predicted samples that are either positive or negative by the Accuracy = (TP + TN) / (TP + FP + FN + TN) formula. Sensitivity is the ratio of correctly defined mixed positives. It is also called the true positive rate and defined as Sensitivity = TP / (TP + FN). Precision is defined as the ratio of accurately recognized failures to all positive predictions, and it equals to Precision = TP / (TP + FP). Error Rate is the measure of 1 – Accuracy. Moreover, the widely used measure that integrates the balance between precision and sensitivity is the F measure by F Measure = 2* Sensitivity * Precision / (Sensitivity + Precision) [62, 63].

This paper develops a novel approach in performance evaluation specific to the telecommunication industry customers. The details will be stated in the modeling part.

3. APPLICATION

In the telecom industry, customer churn is very critical for the operators. For companies in the telecom industry, churn rates of prepaid customers are higher than the churn rates of postpaid customers. The reason for this is the density of periodic users mainly in prepaid customers. Therefore, the tendency of prepaid line users to leave the current company, which is critical and problematic for companies in the telecom sector, is analyzed.

Within the scope of this analysis, it is aimed that the algorithms frequently used in the literature will result in the shortest time and with the most accurate performance criteria, and the methods detailed in the previous sections are tried for this. The variables to be included in the modeling were tested for each performance criterion according to different feature selection methods and the highest output rate was determined. Modeling was done using the Rapid Miner software. It is used for statistical modeling, data preprocessing, business analysis, optimization and predictive analysis [64]. Rapid Miner data processing stages are defining the problem, designing the data requirement, data pre-processing, analytical processing on data and visualization of the data [65].

3.1. Data Definition

The study was conducted using a sample of data telecom company in Turkey. Company name is not shared due to confidentiality. "Ethics Committee Approval document no 12.10.2020-2020/17" was issued for the research. In accordance with the law on protection of personal data, the analysis was carried out on the computers of the company in the working environment of the relevant company, only the results were recorded and no customer data was recorded. The data set to be studied includes 16152 records. 1152 of 16152 customer data are customer data allocated. The customer churn rate has been 7% in mid-2019. There are a total of 27 variables in this data set. The variables used, types of variables and their explanations are shown in Table 2.

Table 2
Variables Used in This Study

	Variables	**Data type**
1	Subscription duration	Numeric
2	Last 6 months balance (rest of pay amount)	Numeric
3	Last 6 months average internet usage	Numeric
4	Customer age	Numeric
5	Days after last device change	Numeric
6	Number of complaints in the last 1 month	Numeric
7	Profit for the last 1 month	Numeric
8	Device group	Text
9	Billing province	Text
10	Internet usage for the last 1 month	Numeric
11	Young segment	Numeric
12	Total number of payments in the last 6 months	Numeric
13	Last 1 month cost	Numeric
14	The average number of calls outside, except abroad, for the last 6 months	Numeric
15	Average usage per person for the last 6 months, in terms of TL	Numeric
16	Average dial-out minutes in the last 6 months	Numeric
17	The average number of incoming calls except abroad for the last 6 months	Numeric
18	Average number of base station cell changes in the last 3 months	Numeric
19	Days since last pay	Numeric
20	Average incoming call minutes in the last 6 months	Numeric
21	Incoming voice call minutes, except abroad, on average in the last 6 months	Numeric
22	Customer gender	Text
23	Total number of lines	Numeric
24	Excess of package	Numeric
25	Excess of package, in terms of TL	Numeric
26	Mobil TV usage	Numeric

The 27th variable indicates whether the customer has left or not, and this variable takes the value 1 if the customer is gone and 0 if the customer is not. Since the text values in the data do not work in support vector machines and artificial neural networks algorithms, they have been made numerical.

3.2. Modeling

Four algorithms selected in modeling were studied; Logistic Regression, Artificial Neural Networks, Simple Bayes Classifier and Support Vector Machines. In these algorithms, 80% -20% training test data were used. In the data set, feature selection was weighted according to information gain, gain ratio, Gini index and correlation, and the resulting results were added in order of variables and algorithms were run.

The aim here is to reduce the noise in the data set. In other words, the purpose here is to determine "what is ineffective" in reaching the most accurate result without including variables having little or no effect. The previously performance criteria were compared, and the results were evaluated. In addition to these performance criteria, which are frequently used in the literature, new performance criteria have been added to measure customer churn in the telecom sector.

The reason for developing a new performance criterion is to ensure that high rate of predicting the customers churn, and ensure that the rate of correctly predicting the customers staying and leaving the company in the total model. Most of the studies are based only on the high rate of accuracy, i.e. customer churn rate. Besides, the sensitivity is not enough for the estimation, the rate of "keeping customers in the existing company" should be predicted well too. By doing so, the customer loss can be understood well and the company can offer them focused campaigns. In most of the studies, sensitivity and accuracy rate can be inversely proportional [66]. This study adopts "Accuracy * Sensitivity = the closest value to 1" measure since the closest and highest values of accuracy and precision ratio will give the best result.

4. FINDINGS

The outputs of the algorithms have been analyzed according to their performance outputs. First, 26 variables are listed according to information gain, gain ratio, Gini index and correlation. Accordingly, when the algorithms are run, the results are very close according to the first 6 variables. Differentiation starts after 6 variables. In Table 3, the results of the selected classification methods are given together with the number of variables that achieve the highest accuracy and sensitivity rates.

According to the interpretation of the Table 3, the results analysis runs with fewer variables have provided higher results. This also has enabled the model to run in a shorter time. Especially since large amount of data is used in the telecom sector, the data density determines the working performance and duration of the model. Fewer variables remove the noise of the model

According to the results in the Table 3, although very close values were reached in other algorithms, the highest accuracy and precision multiplication ratio has been obtained in the Artificial Neural Networks algorithm with the 7th variable as "profit for the last 1 month" set in the feature selection. This is an important result of the churn analysis. When the company's profit is high from that customer, then the churn is inevitable. People noticing that there are many packages with lower cost in other service providers, are tend to change the operator immediately. Also the 9th variable "billing province" has been attracted the attention by addressing that the people locating at the same province affects each other in the operator changing decision. In addition, as a result of the analysis, the 11st "young segment" variable says that young people are more in tendency to change the telecom server, the 17th "the average number of incoming calls, except abroad, for the last 6 months" variable emphasizes that the people having frequent phone calls shift to another operator more, and the 22nd variable "customer gender" is the another factor affecting the mobile service provider leaving decision.

Table 3
Highest Performance Outputs

	Logistics Regression	Support Vector Machines	Naive Bayes	Artificial Neural Networks
Information Gain	12nd, 13rd variables Accuracy:%98,30 Sensitivity: %78,24 Multiplication: 0,769	18 th variable Accuracy: %98,30 Sensitivity: %77,82 Multiplication: 0,778	13. variable Accuracy: %82,54 Sensitivity: %85,36 Multiplication: 0,705	7 th, 9th, 11st, 17th, 22nd variables Accuracy: %98,67 Sensitivity: %84,94 Multiplication: **0,849**
Gain Ratio	6th, 7 th variables Accuracy: %98,27 Sensitivity: %77,41 Multiplication: 0,774	6 th, 22nd, 23rd, 24th, 25th, 26th variables Accuracy: %98,27 Sensitivity: %77,41 Multiplication: 0,760	7 th variable Accuracy: %87,93 Sensitivity: %84,52 Multiplication: 0,743	7 th variable Accuracy: %98,67 Sensitivity: %84,94 Multiplication: **0,849**
Correlation	7 th variable Accuracy: %98,30 Sensitivity: %84,94 Multiplication: **0,835**	12 nd variable Accuracy: %98,33 Sensitivity: %78,24 Multiplication: 0,782	25th variable Accuracy: %81,64 Sensitivity: %84,52 Multiplication: 0,690	7 th , 9th variables Accuracy: %98,67 Sensitivity: %84,94 Multiplication: **0,849**
Gini Index	7 th variable Accuracy: %98,27 Sensitivity: %77,41 Multiplication: 0,774	23rd, 24th, 25th, 26th variables Accuracy: %98,27 Sensitivity: %77,41 Multiplication: 0,760	6 th variable Accuracy: %91,33 Sensitivity: %85,36 Multiplication: 0,779	7 th variable Accuracy: %98,67 Sensitivity: %84,94 Multiplication: **0,849**

Furthermore, the gain ratio and Gini index have produced same results in general. Naive Bayes has been the one giving lowest results. Correlation is the feature giving the highest performance measures in all classification algorithms.

The same analysis has conducted with the updated data in 2021, and the same results have been observed again. Therefore, it is clear that COVID-19 pandemics has not changed the churn behavior of the customers.

5. DISCUSSION

5.1. What happened after the COVID-19 pandemics?

During the COVID-19 period, while many countries around the world experienced serious problems in telecom services, companies with strong digital infrastructure successfully overcame this process and increased the number of customers significantly. It was determined that video calls increased by 50 percent, group video calls by 650 percent, and the use of other digital services increased by 50-60 percent [67]. Network occupancy rates have increased significantly, with mobile traffic increasing by 10 percent and constant 60 percent [68].

The effects of the pandemics on the telecom industry in the short and medium term required adaptation to changing customer habits. With the effect of COVID-19, people spent more time at their homes, increased the frequency of video conversations with their loved ones, and continued their education from home by working from home. These new habits have made the telecom industry one of the most important intermediaries connecting people to each other and the world during the struggle against the pandemics. In the short and medium term, COVID-19 has had both good and bad impacts on the industry in terms of changes in customer demands, supply chain and production-based activities, and operational responsibilities [69].

Despite the positive effects of customer behavior on the telecom industry, the operators experienced delays and decreases in "gaining new customers" due to the fact that customers are in a period of avoiding changes. However, there has been a decrease in the interest of customers, whose most basic needs are to stay connected and in communication, in other telecom products [69].

It has been observed that the most important customer data identified in this study that should be taken into account did not change with the COVID-19 outbreak. The analysis has been done in mid-2019 and has been repeated with the updated data in 2021, and almost the same results as in Table 3 have been observed again. The analysis highlights that only the usage amount has been changed, the customer behavior has remained same.

5.2. Comparison with other telecommunication sector customer churn analysis researches

Özdemir et al. [70] uses machine learning classification algorithms (k-Nearest Neighbors, ANN, NB and Random Forests Algorithm) in Python for the churn analysis in a telecom company, and achieves the maximum accuracy rate with ANN. Although the paper applies similar customer data, it does not underline the most important feature in classification. Pamina et al. [71] conducts a similar study in telecom for churn analysis by using the performance measures as accuracy, recall, precision, F-score, etc. with the aim of producing the highest recall value, and resulted in that the Decision Tree model outperforming all other models. Hooda and Mittal [72] present an exposition of data mining techniques for customer churn in telecom sector, and defines possible factors for churn as price, width of service area, service quality, prepaid packages, curiosity and advertisement. They applied Genetic Programming and Fast Fuzzy C-Means in order to find the expected maximum profit measure for the customer churn. Sharm et al. [73] uses logistic regression and SVM with accuracy, sensitivity, specificity and precision performance metrics. No clear factor is available in the study. Eria and Marikannan [74] apply Random Forest and Logistic Regression, identifies that demographics and customer lifestyles are the most important factors. Gaur and Dubey [75] use Logistic Regression, SVM, Random Forest and Gradient boosted tree algorithms, and emphasizes that Gradient boosting is best in among four models. Al-Shboul et al. [76] plans the churn prediction structure by predicting customers who are likely to churn with

2 heuristic approaches as Fast Fuzzy C-Means and Genetic Programming. Brandusoiu and Toderean [77] urge a complicated technique for predicting the churn in the mobile telecommunications sector by using SVM. Gain measure is in comparisons as performance measurement metric. As a result, the findings of this paper comply with the other researches.

The genuine value of this work hinges upon *(i)* creating a comprehensive feature selection framework as a list based on real customer data from telecommunication industry, *(ii)* providing a comparative study by evaluating the findings of before and after pandemics customer attitudes, *(iii)* presenting applications of different artificial intelligent techniques in telecommunication and indicating these techniques' influence on the findings, *(iv)* specifying the similarities and differences of the findings of this study in comparison with the other existing researches.

While the existing literature presents partial telecommunication customer data (see literature review section), this study takes the advantage of having real telecommunication customer data in order to select the features of the customers. This paper examines the literature in detail, and points out that there are studies focusing on some specific features of the customer. However, this paper utilizes a real telecommunication company's in-use customer feature framework to state the real business management insights of the sector.

6. CONCLUSION

Competition is the main problem facing almost all telecommunications industries around the world. Customer churn in telecommunications is defined as the dissatisfaction of the customers who are leaving the company, or the customer shifting to the other companies that offering reasonable prices. This causes a potential loss of revenue / profit to the companies. Moreover, retaining customers has become a difficult task in the industry. Hence, the companies continue to develop new technologies and applications so that they can provide the best services to their customers. It is necessary to identify customers who are likely to leave the company in the near future, because losing them will result in significant profit loss for the company.

Time management is paramount of importance for the telecommunication companies. In the period of aggressive competition, companies are required to create a quick response according to changing needs. Besides, running big data analysis is both tiring and time consuming for the existing system. In this study, a research was carried out to transform big data into effective data. In other words, the result with big data actually gives the same result when it is done with the specific variables that affect the model. One can reach the same results when they first found the main variables affecting the model and compared the results with both the input variables and the effective variables. As a result, we obtained fewer and more effective variables and proceeded with those variables in future studies. Since it is not working with large data, a great saving was achieved in terms of time.

In this study, the machine learning classification techniques in customer churn analysis have been applied to thousands of customer data, an approach as a new performance criteria have been developed specific to the telecommunication market in order to evaluate the performances of the algorithms, and the most important customer features have been selected in both mid-2019 before the pandemics, and in 2021. Accordingly, the profit amount per customer is the most important criteria in customers' decision of leaving the existing mobile service provider. Secondly, the people in the same location affect each other towards shifting an another operator. Next, young people are more in tendency to change the telecom server. Also the people having frequent phone calls shift to another operator more, while the customer gender is the another factor affecting the mobile service provider leaving decision. Additionally, ANN is the algorithm giving the best results with correlation in feature selection.

Although the rising importance of the telecommunication sector, the customer behaviors have been not altered during the COVID-19 pandemics. The mid-2019 analysis

results, and the newly conducted 2021 results are nearly the same. Only the usage amounts in phone calls, internet and the other services have been increased. The customer churn results have remained the same. Besides, the other current telecom customer churn results support the findings of this paper. Furthermore, as is it seen in the discussion part, the existing researches only groups the customers in terms of some variables. However, they do not clearly illustrate what are these features shaping the customer churn. Hence, this paper contributes to the practitioners by presenting the most important factors. In addition, a new approach in performance evaluation have been proposed specific to the telecommunication market with the industry experts' guidance as a theoretical contribution.

The study is limited only in terms of reflecting a telecom company in one country and representing just one company. However, due to the having a huge amount of data, the sample size is big enough to demonstrate the population's general situation. The paper presents a real-life application, and emphasizes which customer features and which performance metrics should be focused on. Future studies may use different data mining methods with changing data and achieve different results with changing customer profile.

Funding

The authors have no received any financial support for the research, authorship or publication of this study.

The Declaration of Conflict of Interest/ Common Interest

No conflict of interest or common interest has been declared by the authors.

Authors' Contribution

The authors contributed equally to the study.

The Declaration of Ethics Committee Approval

Ethics Committee Approval document no 12.10.2020-2020/17 is taken for the study.

The Declaration of Research and Publication Ethics

The authors of the paper declare that they comply with the scientific, ethical and quotation rules of SAUJS in all processes of the paper and that they do not make any falsification on the data collected. In addition, they declare that Sakarya University Journal of Science and its editorial board have no responsibility for any ethical violations that may be encountered, and that this study has not been evaluated in any academic publication environment other than Sakarya University Journal of Science.

REFERENCES

[1] P.J. Nesse, S.W. Svaet, D. Strasunskas, and A.A. Gaivoronski, "Assessment and optimisation of business opportunities for telecom operators in the cloud value network", Transactions on emerging telecommunications technologies, vol.24, no.5, pp. 503-516, 2013.

[2] T.-H. Chou, J.-L. Seng, "Telecommunication e-services orchestration enabling business process management", Transactions on emerging telecommunications technologies, vol.23, no.7, pp. 646-659, 2012.

[3] Y. Atlı, N. Yücel, "Hibrit iletişim teknolojileri", Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, vol.21, no.3, pp. 785-797, 2016.

[4] Deloitte. "Covid-19 sonrası tedarik zincirlerinde kazananlar ve kaybedenler." https://www2.deloitte.com/tr/tr/pages/consumer-business/articles/Covid-19-sonrasi-tedarik-zinciri.html 2021.

[5] PwC. "COVID-19 salgınının telekom sektörü üzerinde olası etkileri".

https://www.pwc.com.tr/covid-19-telekom-sektoru 2020.

[6] S. Tabassum, M.A. Azad, and J. Gama, "Profiling high leverage points for detecting anomalous users in telecom data networks", Annals of Telecommunications, vol.75, no.9-10, pp. 573-581, 2020.

[7] U. T. Şimşek Gürsoy, "Customer churn analysis in telecommunication sector". Istanbul University Journal of the School of Business Administration, vol.39, no.1, pp.35–49, 2010.

[8] García, D. L., Nebot, À., & Vellido, A. Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, 2017, 51(3), 719–774.

[9] A. K. Ahmad, A. Jafar, & K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform". Journal of Big Data, pp. 6-28, 2019.

[10] M. Al-Mashraie, S.H. Chung, H.W. Jeon, "Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach", Computers and Industrial Engineering, vol.144, 106476, 2020.

[11] N. Alboukaey, A. Joukhadar, N. Ghneim, "Dynamic behavior based churn prediction in mobile telecom", Expert Systems with Applications, 162, 2020.

[12] M. Ahmed, H. Afzal, I. Siddiqi, M.F. Amjad, K. Khurshid, "Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry", Neural Computing and Applications, vol. 32, no.8, pp. 3237-3251, 2020.

[13] M. Hemalatha, S. Mahalakshmi, "Customer churns prediction in telecom using adaptive logitboost learning approach", International Journal of Scientific and Technology Research, vol.9 no. 2, pp. 5703-5713, 2020.

[14] D. Kim, "Investor churn analysis in a P2P lending market", Applied Economics, vol. 52 no. 52, pp. 5745-5755, 2020.

[15] J. Kaur, V. Arora, S. Bali, "Influence of technological advances and change in marketing strategies using analytics in retail industry", International Journal of Systems Assurance Engineering and Management, vol.11 no. 5, pp. 953-961, 2020.

[16] M.A. De la Llave Montiel, F. López, "Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid", Papers in Regional Science, vol. 99 no.6, pp. 1643-1665, 2020.

[17] W. Jiang, Y. Luo, Y. Cao, G. Sun, C. Gong, "On the build and application of bank customer churn warning model", International Journal of Computational Science and Engineering, vol.22 no.4, pp. 404-419, 2020.

[18] P. Verma, "Churn prediction for savings bank customers: A machine learning approach", Journal of Statistics Applications and Probability, vol.9 no.3, pp. 535-547, 2020.

[19] S. Höppner, E. Stripling, B. Baesens, S.V. Broucke, T. Verdonck, "Profit driven decision trees for churn prediction", European Journal of Operational Research, vol.284 no.3, pp. 920-933, 2020.

[20] H. Li, D. Wu, G. X. Li, Y. H. Ke, W. J. Liu, Y. H. Zheng, & X. Lin, "Enhancing telco service quality with big data enabled churn analysis: infrastructure, model, and deployment". Journal of Computer Science and Technology, vol.30 no.6, pp.1201–1214, 2015.

[21] W. Hengliang, & W. Zhang, "A customer churn analysis model in e-business environment". International Journal of Digital Content Technology and Its

Applications, vol. 6 no.9, pp.296–302, 2012.

[22] K. Dahiya, "Customer Churn Analysis in Telecom Industry". 4th International Conference on Reliability, Infocom Technologies and Optimization, pp.1–6, 2015.

[23] M. Günay, "Makine öğrenmesi yöntemleri ile kayıp müşteri analizi", 26th Signal Processing and Communications Applications Conference, pp.1–4, 2018.

[24] P. Kisioglu, & Y. I. Topcu, "Applying Bayesian belief network approach to customer churn analysis : A case study on the telecom industry of Turkey". Expert Systems with Applications, vol.38, pp.7151–7157, 2010.

[25] A. Chouiekh, E.H.I. El Haj, "Deep convolutional neural networks for customer churn prediction analysis", International Journal of Cognitive Informatics and Natural Intelligence, vol.14 no.1, pp. 1-16, 2020.

[26] T. Mandhula, S. Pabboju, N. Gugulotu, "Predicting the customer's opinion on amazon products using selective memory architecture-based convolutional neural network", Journal of Supercomputing, vol.76 no.8, pp. 5923-5947, 2020.

[27] A. De Caigny, K. Coussement, K.W. De Bock, S. Lessmann, "Incorporating textual information in customer churn prediction models based on a convolutional neural network", International Journal of Forecasting, vol.36 no.4, pp. 1563-1578, 2020.

[28] F. Napitu, "Twitter opinion mining predicts broadband internet's customer churn rate". IEEE International Conference on Cybernetics and Computational Intelligence, pp.141–146, 2010.

[29] I. Amali, R. Arunkumar, "Particle swarm optimization with kernel support vector machine for churn prediction in telecommunication industry", International Journal of Scientific and Technology Research, vol.9 no.4, pp. 253-257, 2020.

[30] R. Dong, F. Su, S. Yang, & X. Cheng, "Customer Churn Analysis for Telecom Operators Based on SVM". In: Sun S., Chen N., Tian T. (eds) Signal and Information Processing, Networking and Computer, vol.473, pp.327-333. Springer, Singapore. 2018.

[31] N.N.A. Sjarif, M.R.M. Yusof, D.H.-T. Wong, S. Yakob, R. Ibrahim, M.Z. Osman, "A customer Churn prediction using Pearson correlation function and K nearest neighbor algorithm for telecommunication industry", International Journal of Advances in Soft Computing and its Applications, vol.11 no. 2, pp. 46-59, 2019.

[32] X. Long, Y. Wenjing, A. Le, N. Haiying, L. Huang, Q. Luo, & Y. Chen. "Churn analysis of online social network users using data mining techniques", Lecture Notes in Engineering and Computer Science, vol.2195, pp.551-556, 2012.

[33] F. Fessant, J. François, F. Clérot, "Characterizing ADSL customer behaviours by network traffic data-mining", Annals of Telecommunications, vol.62 no.3-4, pp. 350-368, 2007.

[34] V. Mahajan & Misra. "Review of data mining techniques for churn prediction in telecom", Journal of Information and Organizational Sciences, vol.39 no.2, pp.183–197, 2015.

[35] A. Rodan, A. Fayyoumi, H. Faris, J. Alsakran, & O. Al-kadi, "Negative correlation learning for customer churn prediction : a comparison study". The Scientific World Journal, 1-7, 2015.

[36] Y. Gao, G. Zhang, J. Lu, & J. Ma, "A bi-level decision model for customer churn analysis", Computational Intelligence, vol.30 no.3, pp. 583-599, 2014.

[37] Ç.K. Kontacı, O. Çetintürk, G. G. Polat, K. C. Özkısacık & A. A. Salah, "A simulator for generating realistic simulations of telecom customer behaviors", 24th Signal Processing and Communication Application Conference (SIU), pp. 537-540, 2016.

[38] P. Wanchai, "Customer churn analysis : a case study on the telecommunication ındustry of Thailand". 12th International Conference for Internet Technology and Secured Transactions, pp.325–331, 2017.

[39] V. Gülpınar, & D. Altaş, "Customer churn analysis through artificial neural networks in Turkish telecommunications market", International Journal of Economic Perspectives, vol.7 no.4, pp.63–80, 2013.

[40] N. Forhad, S. Hussain, R. M. Rahman, "Churn Analysis : Predicting Churners". Ninth International Conference on Digital Information Management, pp.237–241, 2014.

[41] Y. Zhao, B. Li, X. Li, W. Liu, S. Ren, "Customer Churn Prediction using improved one-class Support Vector Machine", Lecture Notes in Computer Science, Editör: Li X, Wang S, Yang-Dong Z. 3584, Springer, Berlin, 300–306, 2005.

[42] S. Jamil, & M. S. Cs, "Churn comprehension analysis for telecommunication ındustry using ALBA". International Conference on Emerging Technologies, 1–5, 2016.

[43] K. M. Leung, "Naive Bayesian Classifier".https://web.archive.org/web/20160311202321/http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf. 2007.

[44] C. Budak, M. Türk, A. Toprak, "Removal of impulse noise in digital images with Naive Bayes classifier method". Turkish Journal of Electrical Engineering and Computer Science, vol.24 no.4, pp.2717-2729, 2016.

[45] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis & K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction". Simulation Modelling Practice and Theory, vol.55, 1–9, 2015.

[46] D.W. Hosmer & S. Lemeshow, "Applied logistic regressions", RX Sturdivant, John Wiley & Sons, 1996.

[47] StatisticsSolutions, "What is Logistic Regression?"https://www.statisticssolutions.com/what-is-logistic-regression/ 2021.

[48] B. Huang, M. T. Kechadi, & B. Buckley, "Customer churn prediction in telecommunications". Expert Systems with Applications, vol.39 no.1, pp.1414–1425, 2011.

[49] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms".https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 2018.

[50] F. Filiz, "Artificial neural networks". https://medium.com/@fahrettinf/4-1-1-artificial-neural-networks-6257a7a54bb3 2017.

[51] S. Haykin, "Neural Networks and Learning Machines". Pearson Education, New Jersey, 1999.

[52] Y. Miche, P. Bas, A. Lendasse, C. Jutten, O. Simula, "Advantages of Using Feature Selection Techniques on Steganalysis Schemes". F. Sandoval et al. (Eds.) Springer-Verlag Berlin Heidelberg, pp. 606–613, 2007.

[53] J. Brownlee, "Feature Selection to Improve Accuracy and Decrease Training Time". https://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/ 2021.

[54] O. Kaynar, H. Arslan, Y. Görmez, & Y. E. Işık, "Makine öğrenmesi ve öznitelik seçim

yöntemleriyle saldırı tespiti". Bilişim Teknolojileri Dergisi, pp.175–185, 2018.

[55] M. Santini, "Decision Trees: Entropy, Information Gain, Gain Ratio". https://www.slideshare.net/marinasantini1/lecture-4-decision-trees-2-entropy-information-gain-gain-ratio-55241087?from_action=save 2015.

[56] A. G. Karegowda, A. S. Manjunath, M.A. Jayaram, "Comparative study of attribute selection using gain ratio". International Journal of Information Technology and Knowledge and Knowledge Management, vol.2 no.2, pp.271–277, 2010.

[57] Toppr "Calculation of Gaining Ratio". https://www.toppr.com/guides/principles-and-practices-of-accounting/retirement-of-a-partner/calculation-of-gaining-ratio/ 2021.

[58] L. Breiman, "Classification and regression trees", Chapman & Hall/CRC, 1984.

[59] S. Tahsildar, "Gini Index for Decision Trees".https://blog.quantinsti.com/gini-index/ 2019.

[60] F. Kayaalp, M. S. Başarslan, & K. Polat, "TSCBAS: A novel correlation based attribute selection method and application on telecommunications churn analysis". International Conference on Artificial Intelligence and Data Processing, 2019.

[61] S. Glen, "How to Calculate Pearson's Correlation Coefficient". https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/#Pearson 2020.

[62] M. Yıldız & S. Albayrak, "Customer churn prediction in telecommunication", 23nd Signal Processing and Communications Applications Conference, pp.256-259, 2015.

[63] F. Salfner, M. Schieschke, & M. Malek, "Predicting failures of computer systems: A case study for a telecommunication system". 20th International Parallel and Distributed Processing Symposium, 2006.

[64] A. K. Yadav, H. Malik, & S. S. Chandel, "Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India". Renewable and Sustainable Energy Reviews, vol.52, pp.1093–1106, 2015.

[65] S. Dwivedi, P. Kasliwal, & S. Soni, "Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime)". Symposium on Colossal Data Analysis and Networking, 1-8, 2016.

[66] A. Tharwat, "Classification assessment methods". Applied Computing and Informatics.doi:10.1016/j.aci.2018.08.003 2018.

[67] M. Erkan, "Koronavirüs sürecinde telekom operatörleri nasıl çalıştı?" https://www.hurriyet.com.tr/teknoloji/koronavirus-surecinde-telekom-operatorleri-nasil-calisti-41539689 2020.

[68] C. Deegan, "Koronavirüs sürecinde telekom operatörleri nasıl çalıştı?". https://www.hurriyet.com.tr/teknoloji/koronavirus-surecinde-telekom-operatorleri-nasil-calisti-41539689 2020.

[69] KocDigital. "Covid-19 Telekom sektörü etkileri",https://www.kocdigital.com/blog/covid-19-telekom-sektoru-etkileri 2020.

[70] O. Özdemir, M. Batar, A.H. Işık, "Churn Analysis with Machine Learning Classification Algorithms in Python", Lecture Notes on Data Engineering and Communications Technologies, vol.43, pp. 844-852, 2020.

[71] J. Pamina, J. Beschi Raja, S. Sam Peter, S. Soundarya, S. Sathya Bama, M.S. Sruthi, "Inferring machine learning based parameter estimation for telecom churn prediction", Advances in Intelligent Systems and Computing, 1108 AISC, pp. 257-267, 2020.

[72] P. Hooda, P. Mittal, "An exposition of data mining techniques for customer churn in telecom sector", International Journal of Emerging Trends in Engineering Research, vol.7 no.11, pp. 506-511, 2019.

[73] S. Sharm, S. S. Sushasukhanya "Prediction of customer churn in telecom industries", International Journal of Recent Technology and Engineering, vol.8 no.1, pp. 369-372, 2019.

[74] K. Eria, B.P. Marikannan, "Significance-based feature extraction for customer churn prediction data in the telecom sector", Journal of Computational and Theoretical Nanoscience, vol.16 no.8, pp. 3428-3431, 2019.

[75] A. Gaur, R. Dubey, "Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques", International Conference on Advanced Computation and Telecommunication, 8933783, 2018.

[76] B. Al-Shboul, H. Faris, N. Ghatasheh, "Initializing Genetic Programming using Fuzzy Clustering and its Application in Churn Prediction in the Telecom Industry", Malaysian Journal of Computer Science, vol.28 no.3, 213-22, 2015.

[77] I. Brandusoiu, G. Toderean, "Churn Prediction in the Telecommunications Sector Using Support Vector Machines", Annals of the Oradea University, Fascicle of Management and Technological Engineering, vol.1, pp. 19-22, 2013.