BASIC STATISTICS OF CUSTOMER ANALYSIS DATA IN THE TELECOMMUNICATION SECTOR

Ahmet Yiğit Sunal
ahmetyigit.sunal@deu.edu.tr, https://orcid.org/0009-0005-2414-912X,
Dokuz Eylül University/The Graduate School of Natural And Applied Sciences, Statistics, Data Science, İzmir, Türkiye

Özgül Vupa Çilengiroğlu
ozgul.vupa@deu.edu.tr, https://orcid.org/0000-0003-0181-8376,
Dokuz Eylül University/Faculty of Science, Department of Statistics, İzmir, Türkiye

ABSTRACT: *The telecommunications sector has reached a broad customer base with the increased use of smartphones. However, challenges such as customer churn have also emerged in this sector. This study encompasses a statistical analysis to develop strategies to prevent customer churn by analyzing customer behavior in the telecommunications sector. Basic descriptive and inferential statistics were applied to Ghana's telecommunications sector dataset. The dataset includes demographic information, service usage trends, and customer statuses of 1971 customers from four companies.*

The study evaluated factors affecting customer churn, such as age, gender, education level, duration with the company, payment amount, and reasons for choosing the company. The customer churn rate was found to be 52.5%, with churns being particularly concentrated in the 25-34 age range. Additionally, pricing was identified as the most significant reason for churn, and it was observed that churn rates increased among long-term customers (3-4 years and above).

As the primary motivation of the study was to prepare a dataset for customer churn prediction using machine learning methods, the preprocessing, cleaning, and basic statistical analysis phases of the data were detailed. Furthermore, the findings provided a foundation for advanced analyses and strategic decisions aimed at enhancing customer satisfaction and reducing churn. In this context, the research results offer significant contributions toward the development of customer-focused strategies in the telecommunications sector and the more effective use of machine learning algorithms.

KEYWORDS: *Customer Churn, Telecommunications Industry, Data Preprocessing, Descriptive Statistics*

Date of Submission: 12.12.2024 Date of acceptance: 24.12.2024
https://doi.org/10.5281/zenodo.14593448

1. INTRODUCTION

In the literature, the communication network in which physical objects are connected to each other or larger systems is referred to as the “Internet of Things (IoT)”. One of the primary advantages of IoT is its ability to support decision-making processes by providing real-time data flow through sensors and devices [1]. In recent years, systems used to connect to other devices and systems over the internet and exchange data have incorporated products such as sensors, software, and mobile devices [2]. Notably, the accessibility and high usage rates of mobile devices have led to an increase in their numbers. For instance, smartphone usage was approximately 3.7 billion in 2016, is expected to reach 7.2 billion by the end of 2024, and is projected to reach around 7.9 billion by 2027 [3]. Overall, the extensive use of phones, especially smartphones, has presented a significant opportunity for service providers in the telecommunications sector. However, this opportunity has also brought numerous challenges.

Competition has intensified and become more challenging with the emergence of new markets among service providers. Under these conditions, service providers have shifted their focus to “customer behavior.” Among customer behaviors, “churn” stands out due to the rapidly changing conditions in the telecommunications sector. Strategic decisions by telecom companies to prevent customer churn often focus on customer segmentation and personalized services [4]. As a result, the need to attract new customers and the challenges of retaining existing ones have increased for service providers [5]. Moreover, many service providers face significant customer churn issues due to fierce competition, dynamic market conditions, and the constant introduction of new commercial offers [6].

"Churn analysis," which examines customer loss, has revealed that customer churn rates in the telecommunications sector range between 20% and 40% annually [5]. Additionally, retaining existing customers is found to be 5 to 10 times cheaper than acquiring new ones, and reducing churn by 5% can increase profits by 25% to 85% [7]. Faced with these challenges, service providers have started to develop strategies to prevent and retain customers. These strategies include providing better service quality, affordable tariffs, and promotions to subscribers to maintain and expand their market shares and revenue [8].

However, the large number of subscribers makes it difficult to solve and interpret issues using traditional statistical methods. Machine learning methods have been shown to achieve higher success rates in accurately predicting customer behaviors and reducing churn compared to traditional methods [9]. The increasing volume of data has necessitated the frequent use of machine learning methods in recent years [10].

The purpose of this study is to obtain the fundamental descriptive statistics and basic inferential statistics of large-scale customer churn analysis data and to prepare the data for machine learning algorithms to be used in these methods.

2. METHODOLOGY

In this study the method consists of 3 parts. These parts consist of study design, data description and statistical methods.

Design: The steps necessary for conducting customer analysis on large datasets in the telecommunications sector are planned to be examined in six phases. Following the initial phase, referred to as the "preparatory (research)" phase, the subsequent phases are: "data selection", "data preprocessing and cleaning", "data storage", "data mining (statistical methods)" and "data evaluation and results." Within the scope of this study, the first phase of statistical methods to be used in the "data mining" phase (data description, identification of descriptive statistics, and basic hypothesis testing from inferential statistics) will be carried out. This phase aims to facilitate the description, structuring, and necessary inferences of the data, which are among the most critical steps in data analysis.

Data: The dataset used in this study comprises 10 variables (features) and 1971 participants from open-source datasets of companies operating in Ghana's telecommunications sector [11]. This dataset is a subset of the combined Ghana telecommunications industry data collected from open datasets and annual reports from 2013–2017. It can analyze factors influencing customer churn, evaluate customer satisfaction, and identify trends based on demographics and service usage. Detailed information about the variables and their descriptions is provided in Table 1. All variables have been processed into the system as categorical.

Statistical Methods: All independent variables in the study are categorical. Therefore, descriptive statistics for all data have been calculated as frequencies and percentages. Since the status of customer churn is a binary dependent variable, its frequency and percentage have also been provided. Cross-tabulation was conducted between customer status and all other variables, and chi-square association tests were applied. Additionally, to examine the relationships between all variables for later modeling purposes, the Spearman correlation coefficient was used. All analyses were performed using IBM SPSS Statistics 26 at a significance level of alpha=0.05 [12].

***Table 1.** Variables and descriptions of the Ghana telecommunication sector dataset*

Variables	Explanation
Age	Age range of the customers
Gender	Gender of the customers
Education	Education level of the customers
Employment Status	Employment status of the customers

Telecom Company	Telecommunications provider used by the customers
Reason for Choosing the Company	The primary reason customers choose their telecom provider
Duration with the Company	How long the customers have been with their telecom provider
Payment Method	Type of service plan subscribed by the customers
Monthly Payment Amount	Monthly fees paid by the customers for the services
Customer Churn	A binary indicator of whether the customer has left the service

3. FINDINGS

The customer churn rate for 1971 subscribers registered with different companies in Ghana's telecommunications sector was found to be 52.5% (f=1034) for the years 2013–2017. Descriptive statistics, including frequencies and percentages, were calculated for variables under both customer churn and non-churn scenarios. Additionally, chi-square test results and p-values were obtained to determine the relationships between customer churns and the variables (H_1: There is a relationship between variables) (Table 2).

In the telecommunications data, customer churns were most frequent in the 25-34 age group (32.2%), while approximately 40% of churns were observed among transgender individuals within the gender variable. Among education and employment status categories, the lowest percentages were noted for high school graduates (13.1%) and retirees (13.6%). Notably, customer churns were significantly higher among postgraduate students (33.7%) and self-employed individuals (29.2%). On a company basis, one particular company was predominantly preferred (49.1%), and pricing (44.1%) was identified as the primary reason for choosing that company. The majority of customers had been with their current company for 3-4 years (32.1%).

Additionally, postpaid payment methods were found to dominate (42.3%), and a high payment level (Level 7: 32.2%) was observed. Chi-square tests demonstrated that all these variables were statistically associated with customer churn status (all p-values < 0.05). To analyze the relationships between variables themselves, the Spearman correlation coefficient was utilized (Table 3). Based on this, variables (features or factors) that were correlated with each other were identified, and it was decided not to include these variables simultaneously in machine learning algorithms or simpler regression models.

The variables found to be associated with customer churn included "gender," "education," "reason for choosing the company," "duration with the company," and "monthly payment amount" ($p < 0.05$). When examining the correlation values between other independent variables, for instance, the "age" variable was found to be correlated with "gender," "education," "reason for choosing the company," and "monthly payment amounts." This correlation indicates that careful consideration is required when including these variables in the same model, and the problem can be addressed through multivariate analyses (logistic regression, etc.). Finally, graphical representations of the variables related to customer churns are provided (Figure 1).

It was observed through graphical representations that customer churns were significantly higher among transgender individuals, that pricing was a critical factor in the reason for choosing a company, that postgraduate students exhibited high churn rates, and that the majority of churns occurred among subscribers who had been with their company for 3-4 years. Additionally, as expected, customer churn rates increased with higher payment amounts.

In conclusion, the high churn rates among transgender individuals highlight the need to investigate their specific needs in the telecommunications sector. Similarly, the high churns among postgraduate students suggest that targeted promotions and campaigns for this group could effectively reduce their churn rates. Pricing emerged as the most prominent reason for churns, demonstrating that price sensitivity is a significant factor in customer churn. Improvements in this area are necessary to mitigate customer loss effectively.

Table 2. *Descriptive statistics and chi-square values (p-values) for customer churn variables*

Variables	Customer Churn			
	Yes 1034 (52.5)	No 937(47.5)	Total	Chi-Square (p-value)
Age				
18-24	161(15.6)	233(24.9)	394(20.0)	84.23(0.000)
25-34	333(32.2)	144(15.4)	477(24.2)	
35-44	167(16.2)	164(17.5)	331(16.8)	
45-54	190(18.4)	199(21.2)	389(19.7)	
55-64	183(17.7)	197(21.0)	380(19.3)	
Gender				
Female	218(27.8)	235(33.1)	453(30.4)	8.9(0.011)
Male	253(32.3)	242(34.1)	495(33.2)	
Transgender	312(39.8)	232(32.7)	544(36.5)	
Education				
No Education	136(15.4)	166(21.0)	302(18.1)	40.03(0.000)
Primary	151(17.1)	171(21.6)	322(19.3)	
High School	115(13.1)	133(16.8)	248(14.6)	
Bachelor’s Degree	182(20.7)	155(19.6)	337(20.2)	
Postgraduate	297(33.7)	166(21.0)	463(27.7)	
Employment Status				
Student	142(13.7)	168(17.9)	310(15.7)	40.95(0.000)
Unemployed	152(14.7)	152(16.2)	304(15.4)	
Self-Employed	302(29.2)	165(17.6)	467(23.7)	
Part-Time	144(13.9)	132(14.1)	276(14.0)	
Full-Time	153(14.8)	150(16.0)	303(15.4)	
Retired	141(13.6)	170(18.1)	311(15.8)	
Telecom Company				
A	167(16.2)	179(19.1)	346(17.6)	11.87(0.008)
B	187(18.1)	162(17.3)	349(17.7)	
C	508(49.1)	400(42.7)	908(46.1)	
D	172(16.6)	196(20.9)	368(18.7)	
Duration with Company				
<6 months	159(15.4)	168(17.9)	327(16.6)	31.53(0.000)
6-12 months	202(19.5)	200(21.3)	402(20.4)	
1-2 years	163(15.8)	193(20.6)	356(18.1)	
3-4 years	331(32.0)	198(21.1)	539(26.8)	
>5 years	179(17.3)	178(19.0)	357(18.1)	
Reason for Choosing Company				
Coverage	206(27.8)	223(32.3)	429(30.0)	26.88(0.000)
Customer Service	208(28.1)	253(36.7)	461(32.2)	
Pricing	327(44.1)	214(31.0)	541(37.8)	
Payment Method				
Prepaid	306(29.6)	292(31.2)	598(30.3)	9.65(0.008)
Postpaid	437(42.3)	335(35.8)	772(39.2)	
Both	291(28.1)	310(33.1)	601(30.5)	
Monthly Payment Amount				
Level 1	90(7.7)	88(9.4)	168(8.5)	25.55(0.000)
Level 2	91(8.8)	93(9.9)	184(9.3)	
Level 3	81(7.8)	88(9.4)	169(8.6)	
Level 4	183(17.7)	212(22.6)	395(20.0)	
Level 5	139(12.5)	130(13.9)	259(13.1)	
Level 6	137(13.2)	108(11.5)	245(12.4)	
Level 7	333(32.2)	218(23.3)	551(28.0)	

Table 3. *Spearman correlation coefficients between variables*

Customer Churn*	**p-value**	**Spearman**	**Age***	**p-value**	**Spearman**
Gender	0.003	0.077	Gender	0.000	-0.097
Education	0.000	0.145	Education	0.000	-0.088
Reason for C. Company	0.000	0.110	Reason for C. Company	0.000	-0.130
Duration with Company	0.019	0.053	Payment Method	0.049	0.044
Monthly Payment Amount	0.000	0.101	Monthly Payment Amount	0.014	-0.055
Education*	**p-value**	**Spearman**	**Employment Status***	**p-value**	**Spearman**
Reason for C. Company	0.000	0.193	Duration with Company	0.032	-0.048
Duration with Company	0.000	0.145	Payment Method	0.008	0.060
Monthly Payment Amount	0.000	0.132	**Reason for C. Company ***	**p-value**	**Spearman**
			Payment Method	0.000	0.129

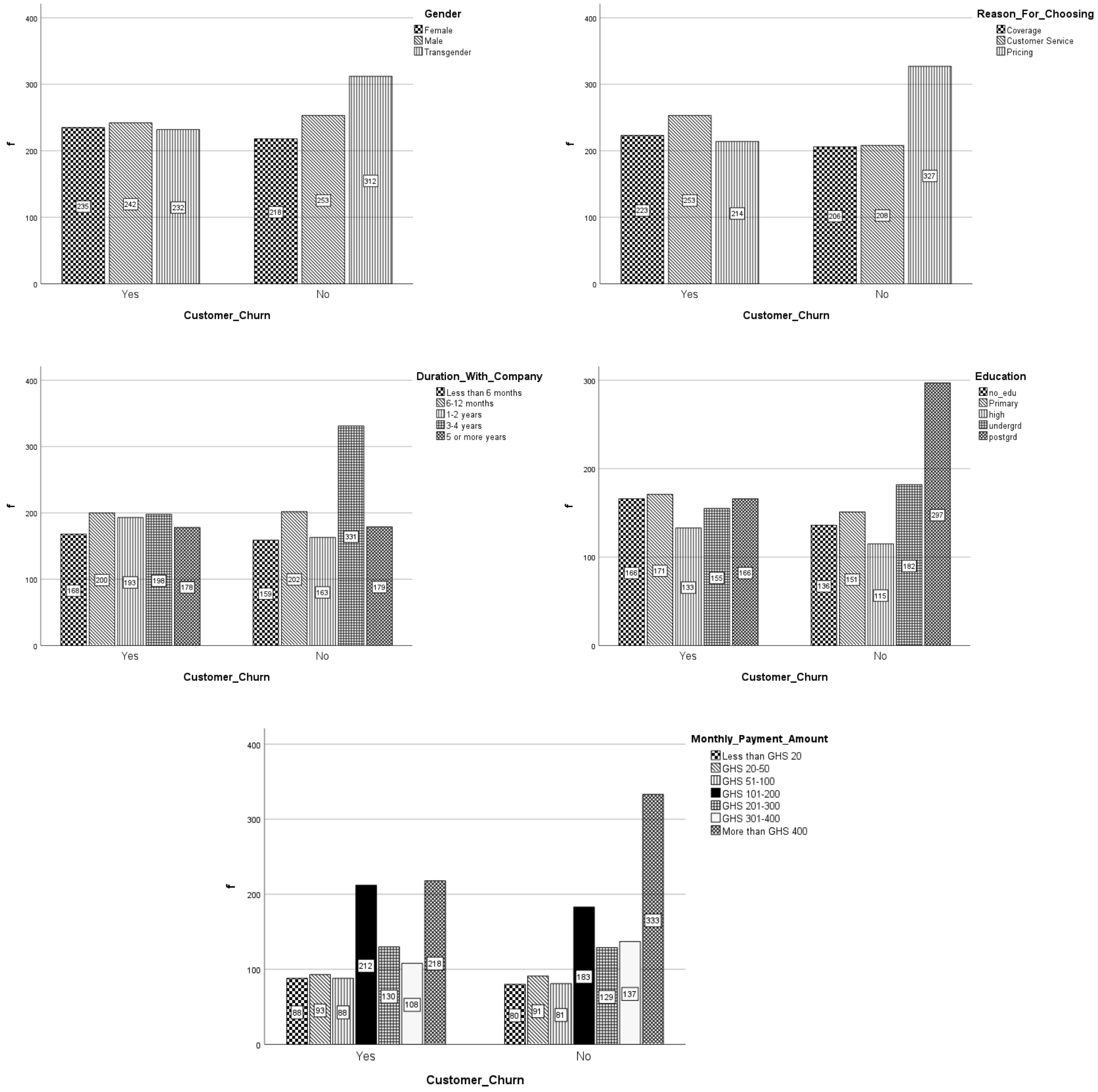

Figure 1. *Customer churn and variables bar charts*

4. DISCUSSION AND CONCLUSION

Today, businesses collect and store large amounts of data. It is evident that when these collected and stored data are utilized correctly, they provide a competitive advantage. The profits, growth rates, and competitiveness of businesses that produce, store, and, most importantly, analyze data accurately are significantly higher. Therefore, under current conditions, statistical analyses, such as those enabled by machine learning, are necessary and inevitable for processing such data. However, to model large datasets effectively, the data must first be processed and organized accurately.

In this study, data from four different companies operating in the telecommunications sector were utilized as preliminary work for machine learning algorithms aimed at predicting customer churn behavior, improving customer relationship management, and ensuring customer retention and continuity.

In the literature, many researchers have explored customer churns in telecommunications data using various machine learning algorithms, resulting in different models. Even within the same dataset, different algorithms have been proposed. For instance, researchers conducted a study on whether 7043 customers from an open-access database had churned or not, employing random forests, support vector machines, and artificial neural networks (ANN) [13]. They found that ANN achieved the highest accuracy of 82%. Similarly, in the literature, ANN's customer churn models have been reported to achieve 86% accuracy [14], 74% accuracy [15], 80% accuracy [16], and 79% accuracy [17].

Likewise, [10] developed models based on k-nearest neighbors, decision trees, random forests, support vector machines, and Naive Bayes algorithms using data from IBM's telecommunications database, which included 7,043 Telco customers. The study concluded that random forests were the best-performing algorithm with an accuracy of 81%. Similarly, researchers found that random forests produced the best model for customer churns, with 79% and 80% accuracy, respectively [18-19]. Researchers, in their study on IBM telecommunications data, chose logistic regression as the most accurate model for customer churns, with 79.8% accuracy among decision trees, logistic regression, support vector machines, artificial neural networks, and Naive Bayes algorithms [20].

These findings highlight that even within the same dataset, different results may emerge, which is a noteworthy outcome. Additionally, high-quality and accurate preprocessing of data is crucial for every study. Proper cleaning and processing of data significantly affect the consistency of results.

Accordingly, in this study, machine learning methods such as decision tree algorithms, Naive Bayes, random forests, artificial neural networks, and logistic regression were chosen to examine customer churns in telecommunications data. In the initial model, variables such as "gender", "education", "reason for choosing the company", "duration with the company" and "monthly payment amount" were included. For subsequent models, variables' relationships were considered to refine the analysis of customer churns. A thorough examination of descriptive statistics should be the first step before building and validating models for each dataset.

REFERENCES

[1]. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems, 29*(7), 1645-1660.

[2]. Wikipedia. (n.d.). Nesnelerin İnterneti. *Wikipedia*. Retrieved from https://tr.wikipedia.org/wiki/Nesnelerin_interneti, Access Date: 01.09.2023

[3]. Statista. (n.d.). Number of smartphone users worldwide from 2016 to 2027. *Statista Research Department*. Retrieved from https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/, Access Date: 01.11.2023

[4]. Verbeke, W., Martens, D., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354-2364.

[5]. Kök, F. (2024). Advanced predictive analytics in customer retention: A study on churn reduction. *Journal of Data Analytics, 10*(2), 105-120.

[6]. Coussement, K., Benoit, D. F., & Van den Poel, D. (2017). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications, 38*(3), 2602–2608.

[7]. Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. Applied Sciences, 11(11), 4742.

[8]. Yeboah-Asiamah, K., Sarpong, O., & Mintah, K. (2018). Strategies for customer retention in telecommunications: The role of service quality and pricing. *International Journal of Marketing Studies, 10*(3), 75-88.

[9]. Ahmed, R., Qamar, U., & Rizwan, M. (2019). A hybrid machine learning framework for customer churn prediction. *Neurocomputing, 349*, 40-53.

[10]. Erdem, Z. U., Çalış, B., & Fırat, S. Ü. (2021). Customer churn prediction analysis in a telecommunication company with machine learning algorithms. *Journal Name*, *32*(3), 496–512.

[11]. Kaggle. https://www.kaggle.com/datasets/freddiej/ghana-telecommunication-data-2023/data, Access Date: 01.10.2023

[12]. IBM SPSS Statistics 29, DEU Bilgi İşlem Daire Başkanlığı, İstatistik Porogramı (2024), SPSS Statistics | Bilgi İşlem Dairesi

[13]. Bilişik, Ö. N., & Sarp, D. T. (2023). Analysis of customer churn in telecommunication industry with machine learning methods. *Düzce University Journal of Science & Technology, 11,* 2185–2208.

[14]. Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019). Customer churn prediction in telecommunication industry using machine learning classifiers. *ACM International Conference Proceeding Series*.

[15]. Raut, N. V. (2020). A study of ensemble machine learning to improve telecommunication customer churn prediction (Master's dissertation, Dublin Business School).

[16]. Hota, L., & Dash, K. (2021). Prediction of customer churn in telecom industry: A machine learning perspective. *Computational Intelligence and Machine Learning, 2*(2), 1–9.

[17]. Makruf, M., Bramantoro, A., Alyamani, H. J., Alesawi, S., & Alturki, R. (2021). Classification methods comparison for customer churn prediction in the telecommunication industry. *International Journal of Advanced and Applied Sciences, 8*(12).

[18]. Tamuka, N., & Sibanda, K. (2020). Real-time customer churn scoring model for the telecommunications industry. *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*.

[19]. Beeharry, Y., & Fokone, R. T. (2022). Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry. *Concurrency and Computation: Practice and Experience, 34*(4).

[20]. Büyükkeçeci, M., & Onur, M. C. (2022). A data mining application in customer churn prediction. DEÜ Faculty of Engineering Journal of Science and Engineering, *24(72)*, 887-900.