RESEARCH ARTICLE OPEN ACCESS

Application of K-Means Clustering in Customer Segmentation

Hoang Thanh Hai*, Mong Thi Nguyet**
*(Thai Nguyen University of Economics and Business Administration, Viet Nam
Email:hoangthanhhai03091988@gmail.com)
** (Thai Nguyen High School, Viet Nam
Email:nguyetmt@tnue.edu.vn)

--************************----------------------------------

Abstract:
Customer segmentation is the process of organizing customers into specific groups based on shared characteristics or behaviours. Segmentation allows marketers to better tailor their marketing efforts to various audience subsets, and improve their products or services. In the telecom industry, an individual might be a client for a few months before churning and trying out a different service. Hence, managing and reducing the churn rate is of crucial importance. In this paper, we use k-means clustering to divide customers into several groups and find out which properties might affect customer churn.

Keywords **— customer segmentation, churn, clustering, k-means.**

--************************----------------------------------

I. INTRODUCTION

In the telecom industry, preventing customer churn is a big concern for service suppliers, since a relatively high percentage of customers leave their providers after a certain time period.It is much more expensive to attract new customers than to retain old ones. On the other hand, focusing on the wrong factors for retaining customers without the intention to leave could be a big waste of time and money. Therefore, it is necessary to identify the clients with a high probability of moving and reduce the influence of factors for it.

There are some types of customer segmentation. The firstis demographic segmentation, which divides customers by age, gender, or marital status. Geographic segmentation is the second type. In this kind of segmentation, we divide clients by their regions. In the third type, we divide customers by their behavior such as purchase history or usage patterns. The last one is psychographic segmentation. The customers are divided based on their lifestyles, interests, and attitudes.

K-means clustering is one of the most widely used unsupervised machine learning algorithms that form data clusters based on the similarity between data instances. Suppose $D = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_N}\}$ is the observed data set. K-means partitions D into K clusters such that the distance between any two clusters is large and the distances between instances inside a cluster are small.

In this paper, we use the Telecom Churn Dataset, which consists of cleaned customer activity data, along with a churn label specifying whether a customer cancelled the subscription. Our main objectives are

- To assess the influences of different variables on churn;
- To cluster customers using k-means;
- To explore each cluster to find out which factor affects client moving.

II. THE METHODOLOGY

The study uses data on 7043 customers. Target variable is “Churn” (Binary – Yes/No). There are 20 features (3 numeric and 17 categorical predictors). The data information is given in Table 1. The features include information about:

- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies;

- Customer account information – how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic information – gender, if they have partners and dependents.

Table 1. Code sheet for variables in the data

	Description	*Codes*
1.	Customer ID	customerID
2.	Client gender (male/female)	gender
3.	Is the client retired (1/0)	SeniorCitizen
4.	Is the client married (Yes/No)	Partner
5.	*How many months a person has been a client of the company (numeric)*	tenure
6.	Is the telephone service connected (Yes/No)	PhoneService
7.	are multiple phone lines connected (Yes, No, No phone service)	MultipleLines
8.	client's Internet service provider (DSL, Fiber optic, No)	InternetService
9.	is the online security service connected (Yes, No, No internet service)	OnlineSecurity
10.	is the online backup service activated (Yes, No, No internet service)	OnlineBackup
11.	does the client have equipment insurance (Yes, No, No internet service)	DeviceProtection
12.	is the technical support service connected (Yes, No, No internet service)	TechSupport
13.	is the streaming TV service connected (Yes, No, No internet service)	StreamingTV
14.	is the streaming cinema service activated (Yes, No, No internet service)	StreamingMovies
15.	type of customer contract (Month-to-month, One year, Two year)	Contract
16.	whether the client uses paperless billing (Yes, No)	PaperlessBilling
17.	payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))	PaymentMethod
18.	*current monthly payment (numeric)*	MonthlyCharges
19.	*The total amount that the client paid for the services for the entire time (numeric)*	TotalCharges
20.	If the client has depentdents (Yes/No)	Dependents
21.	whether there was a churn (Yes or No)	Churn

We use Python for exploratory data analysis, data cleaning, and implementation k-means clustering. There are 11 missing values of TotalCharges. After removing missing values, the final data has 7032 instances.

Mosaic plots and Pearson tests are used to analyse categorical variables and their influences on the target Churn. We use violin plots to assess the impact of numeric variables on the response. Finally, k-means clustering is used to divide customers into different groups. We inspect each group to find out what factors that affected the outcome.

III. Exploratory Data Analysis

3.1 Outcome

The final data has 7032 samples. “Churn” is a binary target with two values: No (5163) and Yes (1869).

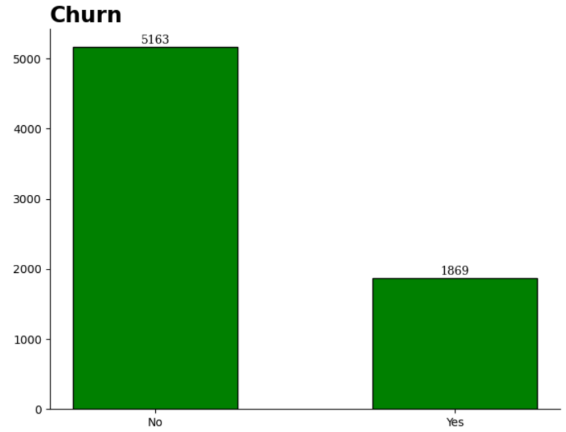

Fig. 1: Bar plot of outcome

3.2 Demographic Variables

There are several demographic attributes in the data set: gender, SeniorCitizen, Partner, Dependents. Using violin plots, we note that gender does not seem to affect on customer churn (figure 2).

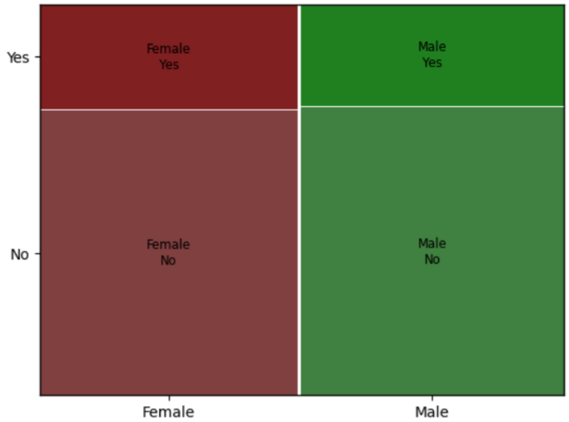

Fig. 2:Mosaic plot of Gender

Individuals with dependents or having a partner are less likely to leave the company, and being a senior citizen or increases the odds of a customer churning (figure 3).

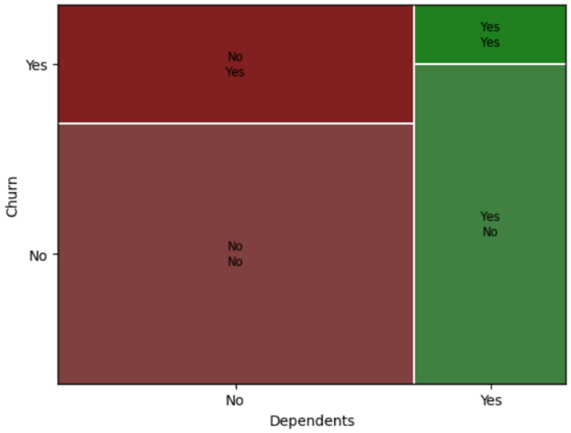

Fig. 3: Mosaic plot of Dependents

We use Pearson tests of independence to examine the above exploratory remark. The results are shown in Table 2. Except Gender, all remaining variables have extremely small p-value, which means that there are clearly dependence between those predictors and outcome Churn.

Table 2. Pearson Tests of Independence *(demographic attributes)*

Variables	*p-value*
Partner	$3.97.10^{-36}$
Gender	0.49
Dependents	$2.02.10^{-42}$
SeniorCitizen	$2.48.10^{-36}$

3.3. Service Variables

There are several service variables in the data set. We also use mosaic plots and Pearson tests to explore the relations between those and the target variable Churn. Most of service variables have influences on Churn (Except for PhoneService and MultipleLines) (Table 3).

Table 3. Pearson Tests of Independence *(Service attributes)*

Variables	*p-value*
InternetService	$5.83.10^{-159}$
OnlineSecurity	$1.40. 10^{-184}$
OnlineBackup	$7.78,10^{-131}$
DeviceProtection	$1.96.10^{-121}$
TechSupport	$7.41. 10^{-180}$
StreamingTV	$1.32.10^{-81}$
StreamingMovies	$5.35.10^{-82}$
Contract	$7.33.10^{-257}$
PaperlessBilling	$8.24.10^{-58}$
PaymentMethos	$1.43,10^{-139}$

Some notices we could see from our mosaic plots are:

- Clients using Fiber optic as the internet service provider have the highest odds of churn;
- Customers without service support and protection (OnlineSecurity, OnlineBackup, DeciveProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling) are more likely to leave the telecom company.
- The shorter the contract is, the more the odds of customer churning are.

3.4 Numeric Variables

We now explore the impact of numeric variables on the outcome by using violin plots. The three numeric predictors are tenure, MonthlyCharges, and TotalCharges.

As we can see in Figure 4, the customers with under one year tenure are much more likely to move.

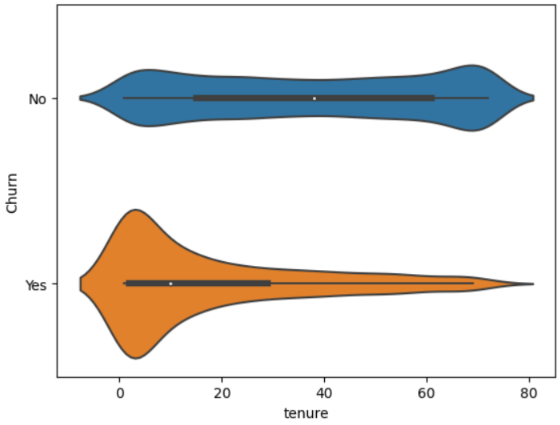

Fig. 4: Violin plot of tenure

It is strange that odds churn in customer group with high monthly charges is higher than that in group with low monthly charges (Figure 5).

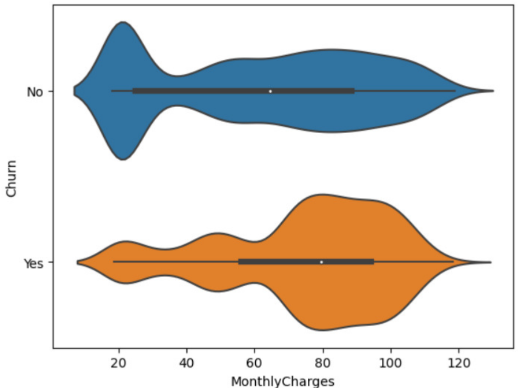

Fig. 5: Violin plot of MonthlyCharges

Therefore, the company should check the reasons why their customer cancelled though they have paid a great amount of monthly charges.

IV. CUSTOMER SEGMENTATION

We use k-means method to segment customers by the months they get telecom services (tenure) and the money they spend monthly (MonthlyCharges). Notice that TotalCharges is determined if we know the corresponding tenure and MonthlyCharges.

The first step is to determine the number of clusters k. Here we use Elbow plot to find out the appropriate value of k. From the plot we choose the value $k = 4$.

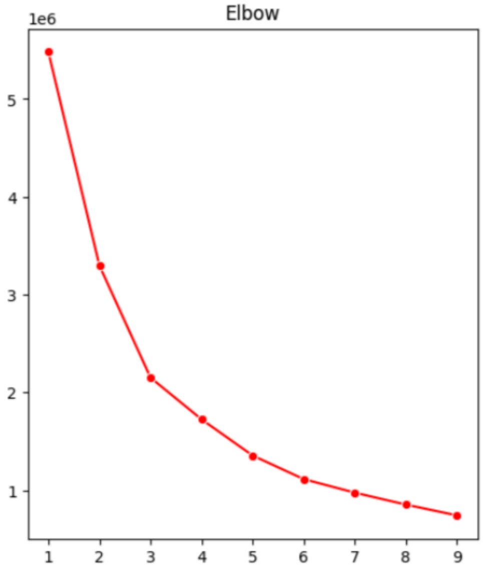

Fig. 6: Elbow plot for choosing the number of clusters.

Fig. 7 shows the four clusters with respect to tenure and monthly charges.

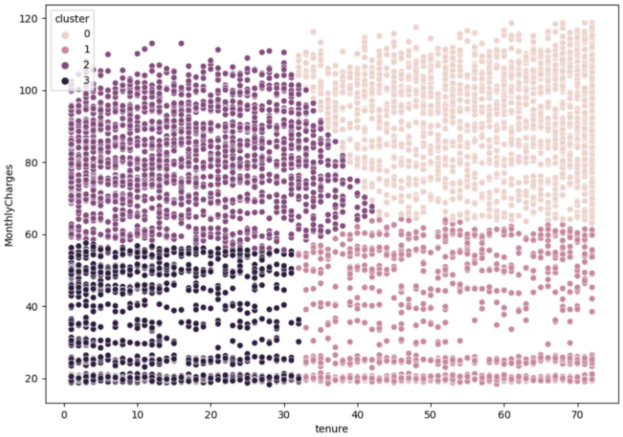

Fig. 7: Four clusters visualization

Among 7043 instances, there are 1869 churned customers. We examine some statistics about this group (Table 4).

Table 4. Statistics churned group

Cluster	*Count*	*Tenure Mean*	*MonthlyCharges Mean*
1	**312**	**52.38**	**98.10**
2	428	5.92	37.79
3	55	49.13	45.54
4	**1074**	**11.20**	**83.65**

We note that there are man clients in cluster 4 (1074/1869), who have been the telecom company clients for nearly one year, and pay an average of 83.65 units monthly, but have moved. Moreover, 312 customers in cluster 1 have paid an average of 98.10 units per month, but have churned after over four years. So, the company should carefully inspect these clusters to find out why they have left.

We now look at some service variables to check if there are any differences between these two groups: yes or no churn. We found that TechSupport could be a reason caused the differences between these two group (Figure 8, 9).

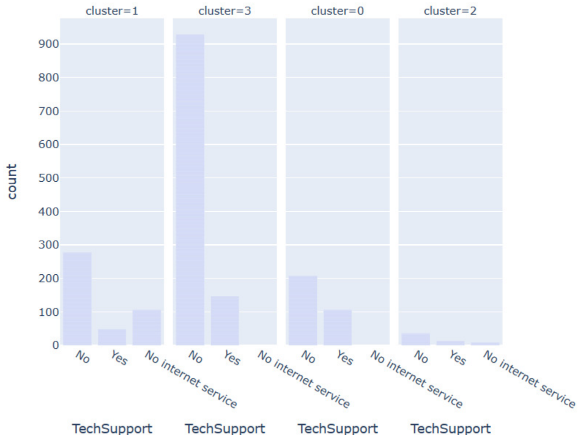

Fig. 8: Bar plot of TechSupport (churn group)

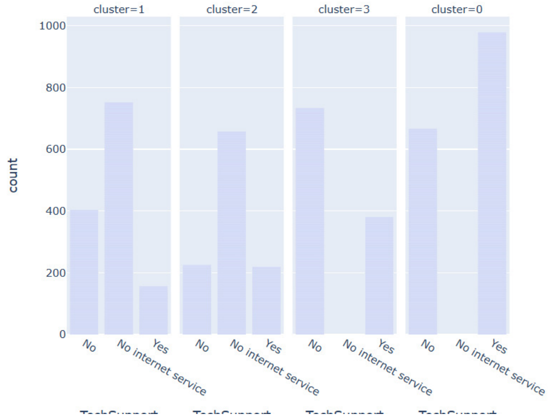

Fig. 9: Bar plot of TechSupport (no churn group)

From the above two plots, we can see that the lack of technical support could be a core reason caused the churning activity. The rate of customers without technical support in group churning clients is far higher than that of no churning users.

IV. CONCLUSIONS

In this paper, we assess the influences of several independent variables on the churn variable. The K-means clustering method is applied to divide the customers of a telecom company into different clusters. We also check to find out which factor could be the reason for churning.

REFERENCES

[1] T. Hastie, R. Tbshirani, J. Friedman, *The Elements of Statiscal Learning*, 2nd ed., Springer, 2001.

[2] A. A. Rodriguez Calderon, *Exploratory data analysis and Customer Segmentation*. [Online]. Available: Exploratory Data Analysis and Customer Segmentatio | Kaggle

[3] A. Shastry, *Customer Retention by Reducing the Churn*. [Online]. Available: Customer Retention by Reducing the Churn | Kaggle