Jurnal Mantik
Volume 4, Number 2, Aug 2020, pp. 1285-1291 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm

Kelvin[1], Cindy[2], Charles[3], Denny Peter Leonardo[4],Yennimar[5]

[1,2,3,4,5]Teknik Informatika,
[1,2,3,4,5]Universitas Prima Indonesia, Jl. Sekip Sei Kambing Medan 20111, Indonesia

E-mail: Tandionokelvin30@gmail.com[1]

ARTICLE INFO

Article history:
Received: 12/01/2020
Revised: 22/07/2020
Accepted: 01/08/2020

Keywords:
Prediction, Churn, Data Mining, FP-Growth

ABSTRACT

Nowadays the competition between companies is increasing. Companies need to predict their customers to find out the level of customer loyalty. One way is to analyze customer data by doing Customer Churn Prediction. In this study the method used is the FP-Growth Algorithm. The FP-Growth algorithm is an algorithm that uses the association rules technique to determine the data that appears most frequently. The data used in this study are secondary data and have 7,403 data from customers. The data has 21 variables. By using a minimum support of 1.2% and confidence at 80%, the associative rules generated are 60. The variable of the type of internet the customer has is strong enough to predict churn. It can be seen that of the 60 associative rules, there are 36 associative rules that have this variable. Testing associative rules on test data yields an accuracy of 71%.

1. Introduction

Currently, new companies are starting to emerge. The emergence of this new company resulted in increased competition. Any company can thrive if it has a sufficient number of subscribers. Having a large number of customers allows optimal revenue to the company's cash. Companies will do various ways to attract these customers to use their services. One way to attract customers is to analyze customer data, which is usually stored in a large number of company databases. These data can be used to analyze which customers are loyal and churn[1] .

The problem faced is how to analyze which customers are loyal or churn. The process of producing information from a collection of data is called data mining. Examples of research that have been carried out by churn analysis are on subscription television company[2], retail company[3], etc. Several data mining methods that have been used to analyze customers are Logistic Regression [1], Decision Tree [4], K-Means [5], Bayesian Belief Network [6], and others [7]. In this study, customer data analysis will be carried out using the FP-Growth algorithm, where the FP-Growth algorithm uses the association rules technique to solve problems.

The FP-Growth algorithm is an algorithm that uses the technique of association rules to determine the most frequently occurring data. Association rules technique is a data mining technique that is used to determine associative rules from item combinations [8]. This technique produces a fairly efficient algorithm. This algorithm is the result of the development of a priori algorithm where the algorithm has a better speed than the a priori algorithm [9]. The FP-Growth algorithm uses a tree concept known as the FP-Tree. Some cases that have been resolved using the FP-Growth algorithm are finding educational promotion strategies [5], recommendation of user product packages[10], and finding association rules in sales transactions[11] .

2. Method

In this study, the work procedures of this study are as follows:

a) Data Selection.
The dataset in this study is secondary data. The dataset contains customer data at a telecommunications company. This dataset is downloaded from the Kaggle website with the website address, namely https://www.kaggle.com/blastchar/telco-customer-churn. The dataset used is 955 KB in size. This dataset

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

has 7,403 customer data and 21 columns. The 21 columns are the variables that will be used as churn predictions. The variables contained in the churn dataset are as follows [12]:

a. customerID: index of the customer.
b. Gender: The gender of the customer. This column has 2 values: make and female.
c. SeniorCitizen: Customers are senior citizens. This column has 2 values, namely 0 and 1.
d. Partners: Customers have a partner. This column has 2 values, namely: Yes and No.
e. Depedents: Customers have dependents. This column has 2 values, namely Yes and No.
f. Tenure: The number of months that the offender uses the company's services.
g. PhoneService: The customer has a telephone service. This column has 2 values, namely Yes and No.
h. MutilpleLines: Customers have multi-channel service. This column has 2 values, namely Yes and No.
i. InternetService: The customer internet service provider. This column has 3 values, namely DSL, Fiber Optic, and No.
j. OnlineSecurity: Customers have online security. This column has 3 values, namely Yes, No and No Internet Service.
k. OnlineBackup:: Customers have online viewing services. This column has 3 values, namely Yes, No and No Internet Service.
l. DeviceProtection: The customer has a device protection service. This column has 3 values, namely Yes, No and No Internet Service.
m. TechSupport: Customers have technical support. This column has 3 values, namely Yes, No and No Internet Service.
n. StreamingTV: Subscribers have a television streaming service. This column has 3 values, namely Yes, No and No Internet Service.
o. StreamingMovies: Subscribers have a movie streaming service. This column has 3 values, namely Yes, No and No Internet Service.
p. Contract: Customer contract terms. This column has 3 values, namely: Month-to-month, One year, Two year
q. PaperlessBilling: Customers have paperless bills. This column has 2 values, namely: Yes and No
r. PaymentMethod: The customer's payment method. This column has 4 values, namely: Electronic check, Mailed check, Bank transfer (automatic), and Credit card (automatic)
s. MonthlyCharges: This is the amount charged to customers each month.
t. TotalCharges: The total amount of services charged to customers.
u. Churn: Category customers churn or not. This column has 2 values, namely Yes, and No.

b) Preprocessing.

After collecting data, the next process is the preprocessing process. Data preprocessing is a process that aims to transform data into a format that is easier and more effective for the user. Some of the data preprocessing methods we use are [13]:

a. Data cleaning is the process of cleaning data that has missing value.
b. Data adjustment is the process of adjusting the amount of data for each target.
c. Data separation, is the process of separating data into two groups, namely train and test.

c) Transformation

Coding is the process of transforming selected data, so that the data is suitable for the data mining process. This process is a creative process and really depends on the type or pattern of information to be searched in the database. In this process, data will be grouped using the One Hot Encoding method.

d) Data mining.

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods [14]. Techniques, methods, or algorithms in data mining vary widely. The research used was FP-Growth. FP-Growth is an algorithm that is included in association rule mining [11]. FP-Growth algorithm is divided into three main steps, namely:

a. Conditional Pattern Base generation stage The Conditional Pattern Base is a subdatabase containing the prefix path and the pattern suffix. The conditional pattern base generation is obtained through the previously built FP-tree.
b. Generation stage for Conditional FP-tree At this stage the support count of each item in each conditional pattern base is added up, then each item that has a greater number of support counts equal to the minimum support count will be generated with a conditional FP-tree.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

c. The frequent itemset search stage, if the Conditional FP-tree is a single path, then the frequent itemset is obtained by combining items for each FPtree conditional. If it is not a single path, then FPgrowth is generated recursively.

e) Interpretation / Evaluation
Information patterns generated from the data mining process need to be displayed in a form that is easily understood by interested parties. At this stage, the knowledge generated by data mining will be released.

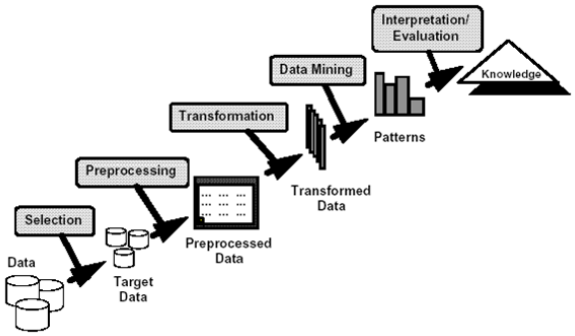

Fig 1. The Process Of Data Mining

3. Result and Discussion

3.1. Result

The object used in this study is customer data at telecommunications companies. Customer data is obtained by downloading on the Kaggle website with the website address https://www.kaggle.com/blastchar/telco-customer-churn. This dataset contains 7,043 customer data. Here are 5 examples of customer data in the dataset.

TABLE 1
EXAMPLE 5 CUSTOMER DATA FROM DATASET

CustomerID	Gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes

TABLE 2
CONTINUED EXAMPLE 5 CUSTOMER DATA FROM DATASET

MultipleLines	InternetService	OnlineSecurity	OnlineBackup	TechSupport	StreamingTV
No phone service	DSL	No	Yes	No	No
No	DSL	Yes	No	No	No
No	DSL	Yes	Yes	No	No
No phone service	DSL	Yes	No	Yes	No
No	Fiber optic	No	No	No	No

TABLE 3
CONTINUED EXAMPLE 5 CUSTOMER DATA FROM DATASET

DeviceProtection	StreamingMovies	Contract	PaperlessBilling	PaymentMethod

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

No	No	Month-to-month	Yes	Electronic check
Yes	No	One year	No	Mailed check
No	No	Month-to-month	Yes	Mailed check
Yes	No	One year	No	Bank transfer (automatic)
No	No	Month-to-month	Yes	Electronic check

TABLE 4
CONTINUED EXAMPLE 5 CUSTOMER DATA FROM DATASET

MonthlyCharges	TotalCharges	Churn
29.85	29.85	No
56.95	1889.5	No
53.85	108.15	Yes
42.3	1840.75	No
70.7	151.65	Yes

The variables used to form association rules are categorical variables. It is necessary to delete some data that are not categorical variables. In this case the variables to be deleted are tenure, MonthlyCharges, TotalCharges and customerID.

The target in this research is the Churn column. The number of data for target No is 5,174 and Yes is 1869. It can be seen that the number of targets is not balanced, so it is necessary to balance the amount of data. The performance of the algorithm will be wrong if the amount of data is not balanced. A method that can be used to balance the amount of data is to remove targets that have a larger amount of data. The amount of data from the results of balancing the data became 3,738 data with a value of No as much as 1,869 data and Yes as many as 1,869 data.

Before forming association rules, it is necessary to divide data into 2 groups, namely train and test. This data sharing aims to analyze whether the association rules generated by the FP-Growth algorithm can be used to predict churn. The distribution of this dataset uses a 4: 3 ratio. By using a 4: 3 ratio, the number of train data is 2,701 and the test data is 1,037.

In this study, the minimum values of support and confidence used for the formation of association rules were 1.2% and 80%. The result of using minimum support and confidence are 60 association rules.

TABLE 5
ASSOCIATION RULES

No	Premises	Conclusion	Support	Confidence	Lift
1	Senior Citizen=No, Contract=Two Year	No	0.15	0.93	1.87
2	Contract=Two year, Partner=Yes	No	0.11	0.83	1.86
3	InternetService=Fiber optic, Contract=Month-to-month, PaymentMethod=Electronic check, OnlineBackup=No	No	0.16	0.93	1.70
4	InternetService=Fiber optic, PaymentMethod=Electronic check, OnlineBackup=No, TechSupport=No	Yes	017	0.92	1.85
5	InternetService=Fiber optic, PaymentMethod=Electronic check, TechSupport=No, Partner=No	Yes	0.12	0.92	1.85
6	InternetService=Fiber optic, Contract=Month-to-month, PaymentMethod=Electronic check, Partner=No	Yes	0.13	0.85	1.71
7	InternetService=Fiber optic, Contract=Month-to-month, PaymentMethod=Electronic check, TechSupport=No	Yes	0.19	0.85	1.70

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

8	PaymentMethod=Electronic check, OnlineBackup=No, TechSupport=No, OnlineSecurity=No	Yes	0.16	0.85	1.70
9	PaymentMethod=Electronic check, OnlineBackup=No, OnlineSecurity=No, PaperlessBilling=Yes	Yes	0.15	0.85	1.69
10	InternetService=Fiber optic, PaymentMethod=Electronic check, OnlineBackup=No, PaperlessBilling=Yes	Yes	0.13	0.84	1.69

TABLE 6
CONTINUED ASSOCIATION RULES

No	Premises	Conclusion	Support	Confidence	Lift
11	InternetService=Fiber optic, Contract=Month-to-month, PaymentMethod=Electronic check, OnlineSecurity=No	Yes	0.19	0.84	1.69
12	PaymentMethod=Electronic check, OnlineBackup=No, TechSupport=No, PaperlessBilling=Yes	Yes	0.15	0.84	1.69
…	…	…	…	…	…
49	InternetService=Fiber optic, TechSupport=No, Partner=No, DeviceProtection=No	Yes	0.14	0.81	1.62
50	Contract=Month-to-month, OnlineSecurity=No, PaperlessBilling=Yes, MultipleLines=Yes	Yes	0.14	0.81	1.62
52	InternetService=Fiber optic , Contract=Month-to-month, PaperlessBilling=Yes, DeviceProtection=No	Yes	0.18	0.81	1.62
53	OnlineBackup=No, TechSupport=No, Partner=No, PaperlessBilling=Yes	Yes	0.15	0.81	1.62
54	InternetService=Fiber optic, Contract=Month-to-month, Partner=No, PaperlessBilling=Yes	Yes	0.16	0.81	1.62
55	InternetService=Fiber optic, OnlineBackup=No , OnlineSecurity=No, PaperlessBilling=Yes	Yes	0.17	0.81	1.62
56	Contract=Month-to-month, TechSupport=No, OnlineSecurity=No, StreamingMovies=Yes	Yes	0.13	0.81	1.62
57	Contract=Month-to-month, OnlineSecurity=No ,Dependents=No, MultipleLines=Yes	Yes	0.14	0.8	1.59
58	InternetService=Fiber optic, TechSupport=No, PaperlessBilling=Yes, DeviceProtection=No	Yes	0.17	0.8	1.59
59	InternetService=Fiber optic, TechSupport=No, Partner=No, PaperlessBilling=Yes	Yes	0.15	0.8	1.59
60	{}	No	0.4	0.4	1

3.2. Discussion

To test the accuracy of the association rules against the test data, confusion matrix method will be used. Here's the confusion matrix from testing association rules:

TABLE 7
CONFUSSION MATRIX

	Prediction +	Prediction -
Actual +	402	187
Actual -	112	336

The process of forming association rules on this data is quite fast, which takes about 3.08 seconds. In the formation of the FP-tree, the time required is 1.72 seconds. It can be seen that FP-Growth has a lot of speed in forming associative rules for variables.

The resulting accuracy from testing the association rules of the test data is 71.17%. Accuracy is calculated using the following methods:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$Accuracy = \frac{402 + 336}{402 + 336 + 112 + 187} \times 100\%$$

$$Accuracy = 71.17\%$$

From the association rules formed using the FP-Growth algorithm, the InternetService variable has a strong enough influence on this churn prediction. This can be seen where from the 60 association rules that are formed, there are 36 rules that have the InternetService attribute.

By knowing that the Internet Service variable has a large enough influence in predicting churn, the company can do better promotions to clients with the Internet Service that has the potential to churn.

4. Conclusion

After conducting this research, the following conclusions can be drawn:

a) FP Growth algorithm can predict the churn problem.
b) The higher the minimum support and confidence, the less association rules are produced.
c) The type of Internet from customers has a considerable influence in predicting churn**.**

5. Reference

[1] T. T. Hanifa and S. Al-faraby, "Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging," *Univ. Telkom*, vol. 4, no. 2, p. 78, 2017.

[2] N. Suryana, "Prediksi Churn Dan Segmentasi Pelanggan Tv Berlangganan (Studi Kasus Transvision Jawa Barat)," vol. 11, no. 2, pp. 185–192, 2016.

[3] N. W. Wardani, G. R. Dantes, and G. Indrawan, "Prediksi Customer Churn dengan Algoritma Decision Tree C4.5 Berdasarkan Segmentasi Pelanggan untuk Mempertahankan Pelanggan pada Perusahaan Retail," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 1, no. 1, pp. 16–24, 2018, doi: 10.31598/jurnalresistor.v1i1.219.

[4] R. Govindaraju, T. Simatupang, and T. A. Samadhi, "Perancangan Sistem Prediksi Churn Pelanggan Pt. Telekomunikasi Seluler Dengan Memanfaatkan Proses Data Mining," *J. Inform.*, vol. 9, no. 1, 2009, doi: 10.9744/informatika.9.1.33-42.

[5] M. P. Syamala, "Analisis Prediksi Churn Dan Segmentasi Pelanggan Speedy Retail Daerah Operasional Bandung Menggunakan Algoritma Decision Tree Dan K-Means," pp. 32–37, 2013.

[6] W. Suharso and A. Djunaidy, "Analisis Customer Churn Menggunakan Bayesian Belief Network (Studi Kasus: Perusahaan Layanan Internet)," *Sisfo*, vol. 4, no. 5, pp. 323–335, 2013, doi: 10.24089/j.sisfo.2013.09.003.

[7] Yulianti, "Metode Data mining Untuk prediksi Churn Pelanggan," *J. ICT Akad. Telkom Jakarta*, vol. 17, no. May, 2018.

[8] A. Ikhwan, D. Nofriansyah, and Sriani, "Penerapan Data Mining dengan Algoritma Fp-Growth untuk Mendukung Strategi Promosi Pendidikan (Studi Kasus Kampus STMIK Triguna Dharma)," *Saintikom*, vol. 14, no. 3, pp. 211–226, 2015.

[9] R. M. Anggraeni, "Perbandingan Algoritma Apriori dan Algoritma FP-Growth untuk Perekomendasi Pada Transaksi Peminjaman Buku di Perpustakaan Universitas Dian Nuswantoro," *Tek. Inform.*, pp. 1–6, 2014.

[10] A. Abdullah, "Rekomendasi Paket Produk Guna Meningkatkan Penjualan Dengan Metode FP-Growth," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, p. 21, 2018, doi: 10.23917/khif.v4i1.5794.

[11] P. Soepomo, "PENGGUNAAN ALGORITMA FP-GROWTH UNTUK MENEMUKAN ATURAN ASOSIASI PADA DATA TRANSAKSI PENJUALAN OBAT DI APOTEK (Studi Kasus : APOTEK UAD)," vol. 2, no. 3, pp. 130–139, 2014, doi: 10.12928/jstie.v2i3.2883.

[12] "Telco Customer Churn | Kaggle." [Online]. Available: https://www.kaggle.com/blastchar/telco-customer-churn. [Accessed: 04-Sep-2020].

[13] R. R. Rerung, "Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk," *J. Teknol. Rekayasa*, vol. 3, no. 1, p. 89, 2018, doi: 10.31544/jtera.v3.i1.2018.89-98.

[14] Nurdin and D. Astika, "Penerapan Data Mining Untuk Menganalisis Penjualan Barang Dengan Pada Supermarket Sejahtera Lhokseumawe," vol. 6, no. 1, pp. 134–155, 2015, doi: 10.29103/TECHSI.V7I1.184.

[15] Sihombing, Oloan, Niskarto Zendrato, Yonata Laia, Marlince Nababan, Delima Sitanggang, Windania Purba, Diarmansyah Batubara, Siti Aisyah, Evta Indra, and Saut Siregar. "Smart home design for electronic devices monitoring based wireless gateway network using cisco packet tracer." *JPhCS* 1007, no. 1 (2018): 012021.

[16] Siregar, S. D., Banjarnahor, J., Dharshinni, N. P., & Tamba, S. P. (2019, July). Understanding group signature methods in making digital signatures to maintain the validity of messages. In *Journal of Physics: Conference Series* (Vol. 1230, No. 1, p. 012072). IOP Publishing.

[17] Banjarnahor, J., Siregar, S. D., Sihombing, O., Turnip, M., Purba, W., Aisyah, S., & Banjarnahor, J. (2019, July). Audio steganography applications using auditory features watermarking. In *Journal of Physics: Conference Series* (Vol. 1230, No. 1, p. 012073). IOP Publishing.