JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

CLUSTERING ROTATIONAL CHURN OF TELECOMMUNICATIONS CUSTOMERS USING A DATA-CENTRICAI APPROACH

Widang Muttaqin[1*], Maya Silvi Lydia[2], Fahmi[3]
[1]Sains Data dan Kecerdasan Buatan, Universitas Sumatera Utara
[2]Ilmu Komputer, Universitas Sumatera Utara
[3]Teknik Elektro, Universitas Sumatera Utara
email: *widangmuttaqin@gmail.com

Abstract: In the current era of very fast technological development, customer churn is a serious challenge, especially in the competitive telecommunications industry. Churn refers to customers who stop using a service or move to another provider, and can be categorized into three types: Active Churn, Passive Churn, and Rotational Churn. Rotational Churn, which is difficult to predict be- cause the reasons for stopping are unclear, is the main focus of this research. This research aims to group Rotational Churn customers using a Data-Centric AI approach. This approach emphasizes improving data quality through Confident Learning and Synthetic Data before being applied to the K-Means clustering algorithm. The data used in this research is customer churn data from one telecommunications company during 2023. The research results show that customer grouping using the K-Means algorithm can provide deep insight into the characteristics of customer churn. The application of Data-Centric AI is proven to be able to increase the accuracy of clustering models, which ultimately helps compa- nies optimize programs and services to minimize churn and retain customers.

Keywords: data-centric AI; clustering; K-means

Abstrak: Dalam era perkembangan teknologi yang sangat pesat saat ini, churn pelanggan menjadi tantangan serius, terutama dalam industri telekomunikasi yang sangat kompetitif. Churn mengacu pada pelanggan yang berhenti menggunakan layanan atau beralih ke penyedia lain, dan dapat dikategorikan menjadi tiga jenis: Churn Aktif, Churn Pasif, dan Churn Rotasional. Churn Rotasional, yang sulit diprediksi karena alasan penghentian layanan tidak jelas, menjadi fokus utama penelitian ini. Penelitian ini bertujuan untuk mengelompokkan pelanggan Churn Rotasional menggunakan pendekatan Data-Centric AI. Pendekatan ini menekankan pada peningkatan kualitas data melalui Confident Learning dan Synthetic Data sebelum diterapkan ke algoritma K-Means clustering. Data yang digunakan dalam penelitian ini adalah data churn pelanggan dari satu perusahaan telekomunikasi selama tahun 2023. Hasil penelitian menunjukkan bahwa pengelompokan pelanggan menggunakan algoritma K-Means dapat memberikan wawasan mendalam tentang karakteristik churn pelanggan. Penerapan Data-Centric AI terbukti mampu meningkatkan akurasi model klastering, yang pada akhirnya membantu perusahaan mengoptimalkan program dan layanan untuk meminimalkan churn serta mempertahankan pelanggan.

Kata kunci: data-Centric AI; klasterisasi; K-means

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Customers are individuals or groups who transact by purchasing products or services from a business. Customers are crucial to all business sectors, as businesses cannot grow or survive without them [1].

Every business competes with others to add or maintain customers by aggressively marketing and promoting products or creating new products and unique experiences that customers like. Companies also pay close attention to customer interactions to learn about customer behavior. There are two major categories in customer grouping: internal customers and external customers. Internal customers originate from within the company or organization or from groups that cooperate or make agreements to generate mutual benefits, while external customers come from all societal levels, such as housewives, students, traders, and others [2].

In all dynamic and competitive business segments, retaining customers is considered a valuable asset for companies to keep their business running [3]. In a highly competitive market, customers have many choices for service providers; switching or even stopping service usage may be easy. This phenomenon is known as "customer churn."[4].

Churn, or customer attrition, is a severe and challenging issue that can affect businesses and industries, particularly the rapidly evolving and competitive telecommunications sector [5].

Churn impacts business performance, such as decreasing sales due to customers using products for a short time or customer dissatisfaction with a purchased product. This allows competitors to acquire churned customers. The higher the churn rate, the greater the risk to the company, whether related to profit or the company's reputation.

Research by [6] proposed a hybrid model combining clustering and classification techniques based on ensemble methods using various clustering techniques such as K-Means, K-Medoids, and random clustering. Data from GitHub and Bigml were used in this study, with results showing the suggested model achieved the highest prediction accuracy of 94.7% and 92.43% on GitHub and Bigml datasets, respectively. Clustering to create a model performs better than the most sophisticated churn prediction models. Another study by [7] stated that clustering generally involves partitioning a dataset consisting of n points embedded in m- dimensional space into k unique clusters so that data points in the same cluster are more comparable to each other than to data points in other clusters. The researcher presented a distance metric that works well for mixed numeric and categorical data sets, with distance measurement accuracy demonstrated by findings from various mixed data sets handled by the K-Means algorithm.

Several studies mentioned above demonstrate various methods used in the clustering process to group churn customers. A good model can be obtained from the proper method selection, hyperparameter tuning, and optimization during the training process. The above studies also used models based on the objective of generating the best model for a specific data set, an approach known as Model-Centric AI. Another approach is based on the concept of Data-Centric AI. Data-Centric AI is an artificial intelligence concept that focuses on data, supporting a fundamental shift from model refinement to ensuring data quality and reliability [8].

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Although manual exploratory data analysis is a crucial initial step in understanding and refining any data set, data-centric AI uses AI technology to detect and address issues that often plague real-world data sets more systematically.

Model-Centric AI strives to provide the best model for a specific data set, while Data-Centric AI aims to generate the best data set systematically and algorithmically to support a particular Machine Learning model. The Data-Centric AI workflow includes data exploration, training a baseline Machine Learning model with a proper data set, and data utilization to enhance the data set [9].

The Data-Centric approach to artificial intelligence focuses on refining current data sets. This requires efforts to improve the quality, accuracy, and usability of the data used in analysis or modeling. Confident Learning is one strategy used in this approach, where a machine learning model is trained to detect inaccurate labels in the data set, which are then corrected or removed [10].

METHOD

General architecture is a series or scheme of the design of a program or system to be built. The general architecture is divided into three main parts: Data Collecting, Data Preprocessing, and Data Modelling.

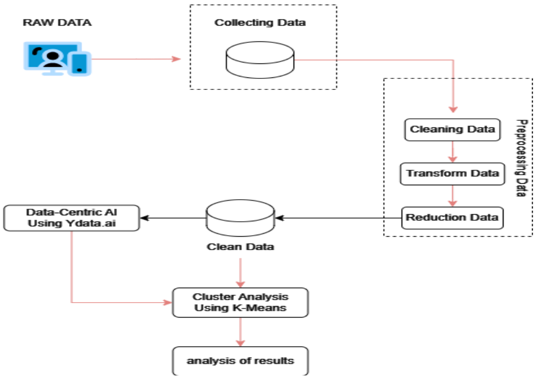


Figure 1. General Architecture

The table below provides a brief overview of the data used in this study.

Table 1. Summary of the data characteristics

Jumlah Baris	109.701
Jumlah Kolom	21
Missing Value	67.786
Tipe Data Kate-gorik	57.14%
Tipe Data Numer- ik	28.57%
Tipe Data String	9.52%

To begin the cleaning process, the author first displays the information of each column as shown in the image 4.

```
RangeIndex: 109701 entries, 0 to 109700
Data columns (total 21 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   a.msisdn              109701 non-null  int64
 1   deactivation_date     109701 non-null  object
 2   deactivation_month    109701 non-null  int64
 3   bc                    109701 non-null  int64
 4   city_hlr              109701 non-null  object
 5   region_hlr            109701 non-null  object
 6   area_hlr              109701 non-null  object
 7   region_lacci          102393 non-null  object
 8   area_lacci            102393 non-null  object
 9   branch_lacci          102393 non-null  object
 10  los                   109701 non-null  int64
 11  los_type              109701 non-null  object
 12  tot_bill_amount       109701 non-null  int64
 13  rev_segment_tot_bill  109701 non-null  object
 14  sub_total             109701 non-null  int64
 15  rev_segment_sub_total 109701 non-null  object
 16  offer_name            109701 non-null  object
 17  offer_map             109701 non-null  object
 18  flag                  109701 non-null  object
 19  channel               63839 non-null   object
 20  churn_type            109701 non-null  object
```

Figure 2 Column Information

From the image above, the author gains deeper insights into the dataset used. The image shows there are 67,786 missing cells and 611 duplicate rows,

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

which accounts for 0.6% of the total data. Additionally, the memory usage is 17.6 MB.

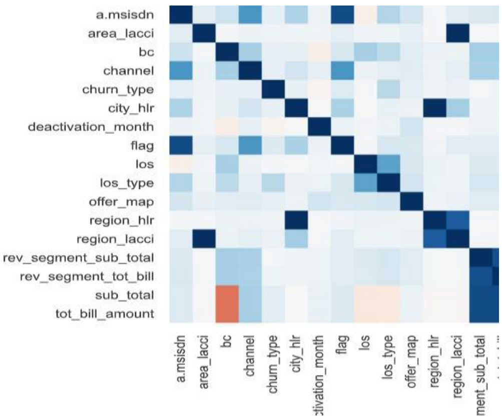


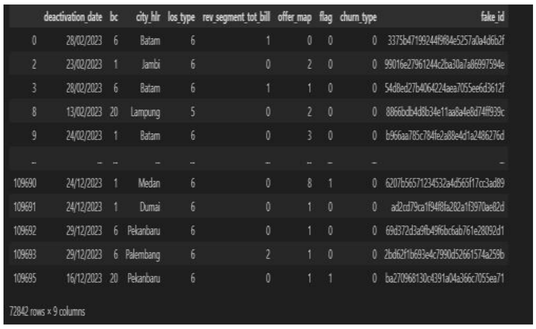
Figure 3. Heatmap

Transform Data

This process involves transforming the previously cleaned data, such as data analysis, data mining, and machine learning. Data transformation is carried out to meet assumptions and analysis or to optimize the performance of the model to be created. df_new.los_type.replace({'00. Unknown':0,'01. 0 - 3M':1,'02. 3 - 6M':2,'03. 6 - 12M':3,'04. 1 - 3Y':4,'05. 3 - 7Y':5,'06. > 7Y':6}, inplace=True) df_new.flag.replace({'PSB':0,'P TP':1}, inplace=True) df_new.churn_type.replace({'Involuntary':0},inplace=True) df_new.offer_map.replace({'Halo Unlimited':0,'Halo+':1,'Halo Kick':2,'Flash':3,'Halo Orbit':4,'Halo Hybrid':5,'Halo Play':6,'Halo Fit':7,'Non Bundle':8}, inplace=True) df_new.rev_segment_tot_bill.replace({'01. <100K':0,'02. 100K - 300K':1,'03. 300K - 1000K':2,'04. >= 1000K':3}, inplace=True).

The program code above shows the process of changing the contents of several columns, such as the columns los_type, flag, churn_type, offer_map, and rev_segment_tot_bill. These columns will later serve as references for the author in performing the clustering process on the dataset.df_new.deactivation_date =pd.to_datetime(df_new.deactivation_date).

The author changed the data format in the deactivation_date column to datetime to facilitate the subsequent processes. The code snippet can be seen above. The results of this code can be observed in the image 4.



	deactivation_date	bc	city_hlr	los_type	rev_segment_tot_bill	offer_map	flag	churn_type	fake_id
0	28/02/2023	6	Batam	6	1	0	0	0	3375b47199244f9f84e5257a0a4d6b2f
2	23/02/2023	1	Jambi	6	0	2	0	0	99016e27961244c2ba30a7a86997594e
3	28/02/2023	6	Batam	6	1	1	0	0	54d8ed27b4064224aea7055ee6d3612f
8	13/02/2023	20	Lampung	5	0	2	0	0	8866bdb4d8b34e11aa8a4e8d74ff939c
9	24/02/2023	1	Batam	6	0	3	0	0	b966aa785c784fe2a88e4d1a2486276d
...	...	...	...	...	...	...	...	...	...
109690	24/12/2023	1	Medan	6	0	8	1	0	6207b56571234532a4d565f17cc3ad89
109691	24/12/2023	1	Dumai	6	0	1	0	0	ad2cd79ca1f94f8fa282a1f3970ae82d
109692	29/12/2023	6	Pekanbaru	6	0	1	0	0	69d372d3a9fb49f6bc6ab761e28092d1
109693	29/12/2023	6	Palembang	6	2	1	0	0	2bd62f1b693e4c7990d52661574a259b
109695	16/12/2023	20	Pekanbaru	6	0	1	1	0	ba270968130c4391a04a366c7055ea71

72842 rows × 9 columns

Figure 4. Result of Duplicate Data Removal

The remaining amount of data after the remove duplicate process is 72,842, down from the previous total of 73,250.

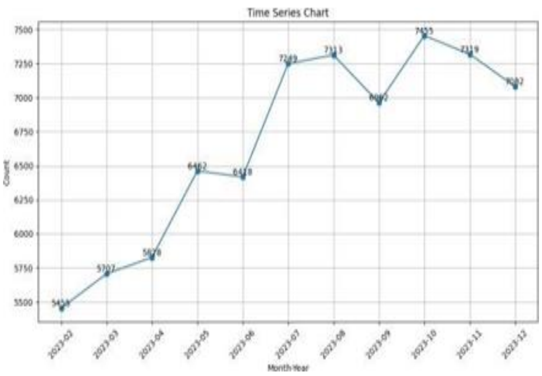


Figure 5. Total Churn per Month

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

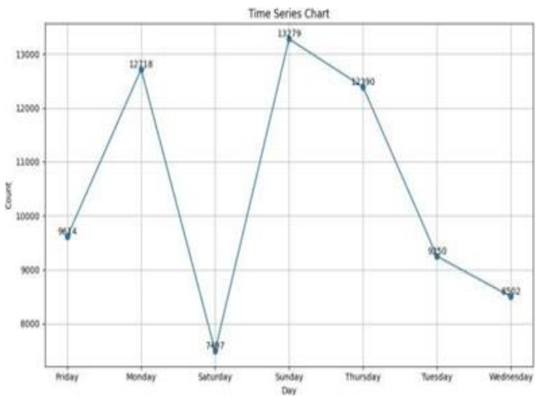


Figure 6 . Total Churn per Day

The image above shows the total churn per month. Churn experienced a significant increase during the June – July period. For daily churn, the increase occurs on Sundays.

Data Modelling

Data modeling is the process of defining the structure of data and its relationships to represent information in a structured and meaningful way. This process involves identifying entities, attributes, and relationships between entities within a specific domain. Data modeling is the initial step in developing information systems, data analysis, or database design. The goal of data modeling is to understand and organize data well so that it can be accessed, managed, and manipulated efficiently.

```
Fitur 1:
  Rata-rata: 3.589012565343926
  Standar Deviasi: 1.4305280481482037
Fitur 2:
  Rata-rata: 0.6022261761906732
  Standar Deviasi: 0.5828025162228823
Fitur 3:
  Rata-rata: 1.0513096371084536
  Standar Deviasi: 1.3249300447913464
Fitur 4:
  Rata-rata: 0.5296478669259734
  Standar Deviasi: 0.4991202299914718
```

Figure 7. Mean and Std Deviation 4 features

RESULTS AND DISCUSSION

Define K Optimun

Determining the number of clusters or the value (k) to be used in the clustering process can be determined by the Elbow method. This method is used to determine the number of clusters or the value (k) in a data set. The Elbow method is relatively easy to understand and implement by looking at the elbow in the inertia graph and using the one with the highest degree elbow as the number of clusters to use.

Clustering Dataset menggunakan K-Means

In this stage, the preprocessed dataset is clustered using the K-Means algorithm with k=5 based on the previous selection process.

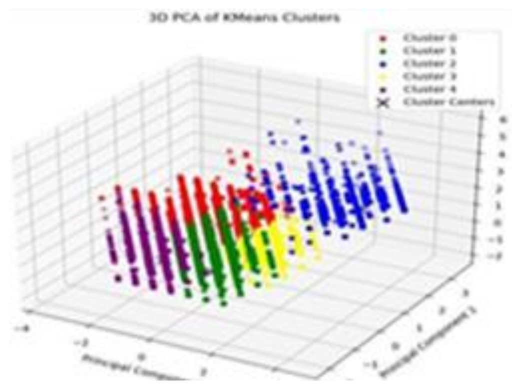


Figure 8. PCA Clustering

Based on the image 8, it can be concluded that the churn data has been divided into 5 clusters based on the total bill amount compared to the loss type.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 773 – 780 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.3304
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 2. Clustering Result for 4 Variable

Clus-ter	Bc	Tot_Bill_Amount	Sub_Total	Fake_Id
0	8.903.474	73.062.230.439	65.821.904.104	15.861
1	1.950.006	162.886.592.872	146.744.778.615	16.862
2	8.168.620	55.770.786.633	50.244.010.262	19.197
3x	9.865.684	35.832.386.947	32.281.491.368	2.375
4	4.910.570	87.329.633.728	78.675.388.260	15.554

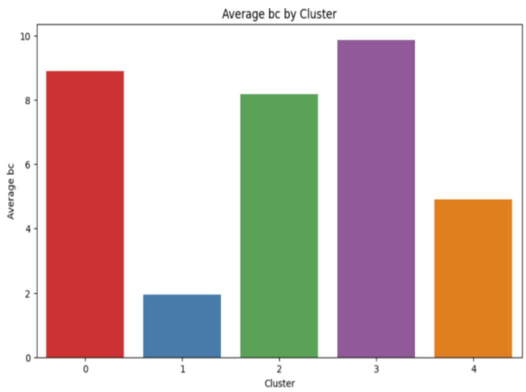


Figure 9. Average bc by Cluster

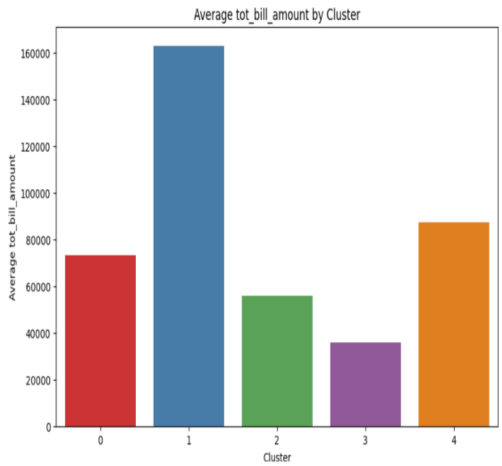


Figure 10. Average tot_bill by Cluster

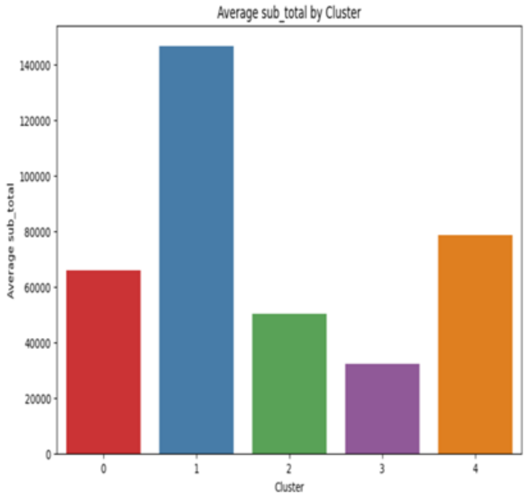


Figure 11. Average sub_total by Cluster

Cluster 0 has an average BC of 8,903, an average total bill amount of 73.06 million, and an average sub_total of 65.82 million, derived from 15,861 fake IDs. Cluster 1, with 16,862 fake IDs, has the highest averages in total bill amount and sub_total, which are 162.89 million and 146.74 million, respectively, with an average BC of 1,950. Cluster 2, with an average BC of 8,168, a total bill amount of 55.77 million, and a sub_total of 50.24 million, has the highest number of fake IDs at 19,197.

Cluster 3, with 2,375 fake IDs, has an average BC of 9,865, an average sub_total of 32.28 million, and a total billing of 35.83 million. The last cluster, Cluster 4, has 15,554 fake IDs, with an average total bill of 87.32 million, an average BC of 4,910, and a sub_total of 78.68 million.

Table 3. Computed Values of Revenue

Cluster	Rev Segment Total Bill <100K	Rev Segment TotalBill 100K - 300K	Rev Segment TotalBill 300K - 1000K	Rev Segment SubTotal Bill <100K	Rev Segment SubTotal Bill 100K - 300K	Rev Segment SubTotal Bill 300K - 1000K	LOSType 2	LOSType 3	LOSType 4	LOSType 5	LOSType 6	BC 1.0	BC 11.0	BC 20.0	Flag 0	Flag 1
Cluster 0	39.00%	60.99%	0.00%	39.83%	60.17%	0.00%	43.77%	35.51%	0.00%	0.00%	0.00%	60.98%	25.24%	6.26%	97.37%	2.63%
Cluster 1	2.49%	91.95%	5.56%	2.89%	94.66%	2.45%	31.84%	23.52%	24.40%	5.19%	0.72%	90.95%	1.60%	0.12%	4.29%	95.71%
Cluster 2	92.76%	6.86%	0.38%	93.43%	6.44%	0.13%	2.44%	2.41%	16.51%	22.61%	52.34%	30.99%	23.45%	30.06%	46.10%	53.89%
Cluster 3	64.60%	35.06%	0.34%	66.57%	33.31%	0.12%	0.00%	0.00%	47.33%	35.14%	17.53%	29.45%	37.74%	17.52%	100.00%	0.00%
Cluster 4	69.46%	30.54%	0.00%	71.75%	28.25%	0.00%	3.53%	5.43%	47.60%	35.51%	7.89%	36.59%	40.17%	13.64%	0.00%	100.00%

The table containing the results of the data analysis for the tested features can be seen.

Table 4. Computed Values of Revenue Segmnetation Based on Clusters by Synthetic Data

Cluster	Rev Segment Total Bl <100K	Rev Segment TotalBill 100K-300K	Rev Segment TotalBill 300K-1000K	Rev Segmen SubTota Bl <100K	Rev Segmen SubTota Bl 100K 300K	Rev Segmen SubTota Bl 300K 1000K	LOSType 2	LOSType 3	LOSType 4	LOSType 5	LOSType 6	BC 1.0	BC 11.0	BC 20.0	Flag 0	Flag 1
	59.80%	39.07%	1.11%	62.40%	37.17%	0.44%	0.00%	0.00%	46.34%	35.22%	18.44%	31.68%	34.01%	17.02%	100.00%	0.00%
	66.03%	33.95%	0.01%	68.64%	31.35%	0.01%	3.64%	5.33%	48.13%	34.34%	8.56%	39.06%	36.71%	13.20%	0.00%	100.00%
	91.49%	8.11%	0.40%	92.78%	6.92%	0.30%	2.87%	2.37%	15.48%	24.63%	50.94%	31.31%	20.87%	29.08%	47.82%	52.18%
	39.10%	60.84%	0.06%	40.52%	59.48%	0.00%	42.71%	34.36%	0.00%	0.00%	0.00%	61.08%	23.89%	6.25%	97.24%	2.76%
	2.52%	89.09%	8.39%	3.07%	92.17%	4.76%	31.27%	23.48%	24.18%	6.07%	0.78%	89.64%	2.03%	1.52%	5.13%	94.87%

In the second dataset, which underwent data synthesis using ydata AI and Generative AI, clustering was performed using K-Means, resulting in 5 clusters. Cluster 0 has the majority of customers with billing segments of <100K (59.80%) and 100 – 300K (39.07%), dominant in BC11 (34.01%) and 01 (31.65%), with all having the PSB flag (100%). Cluster 1 has the majority of customers with BC1 (39.06%) and BC11 (36.71%), with billing <100K (66.03%) and 100 – 300K (33.95%), dominated by two major LoS categories: 1 – 3 years (48.13%) and 3 – 7 years (34.34%), and all with the P2P flag (100%). Cluster 2 has the majority of billing in the <100K category (91.94%), divided into 3 LoS categories: >7 years (50.94%), 3 – 7 years (24.63%), and 1 – 3 years (15.48%), with flags evenly distributed (PSB 47.82% and P2P 52.18%).

Cluster 3 consists of mid-range billing segments of <100K (39.10%) and 100 – 300K (60.84%), divided into two BC categories (61.08% BC1 and 23.89% BC11), spread across 2 LoS categories: 3 – 6 months (42.71%) and 6 – 12 months (34.36%), and dominated by the PSB flag (97.24%). The last cluster, Cluster 4, has high customer billing categories (89.09% in 100 – 300K and 11.91% in 300 – 1000K), with the majority in BC01 (89.64%), and dominated by the P2P flag (94.87%).

CONCLUSION

From the testing, special attention needs to be given to clusters 2 and 4 in the original data. Cluster 2 has significant variation in total billing and loss type, with some customers having high billing amounts but low subtotals. Cluster 4 shows higher variation in total billing and a clearer positive relationship between total bill amount and subtotal. Customer management and offer strategies could focus on maximizing the value of customers with high billing amounts. The results of this study provide insights into the use of both original and synthetic data in the clustering process of customer churn in the telecommunications sector. The original data presents different patterns and visualizations compared to the synthetic data, even when using the same data features. These findings can help companies create targeted programs aimed at specific clusters to reduce churn rates, thereby positively impacting the company's revenue growth.

BIBLIOGRAPHY

[1] "Customer: Definition and How to Study Their Behavior for Marketing." Accessed: Nov. 13, 2023. [Online]. Available: https://www.investopedia.com/terms/c/customer.asp

[2] B. A. Lukas and I. Maignan, "Striving for quality: The key role of internal and external customers," J. Mark. Manag., vol.

1, no. 2, pp. 175–187, 1996, doi: 10.1007/bf00128689.

[3] K. Coussement, S. Lessmann, and G. Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry, vol. 95. Elsevier B.V., 2017. doi: 10.1016/j.dss.2016.11.007.

[4] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture," Expert Syst. Appl., vol. 85, pp. 204–220, 2017, doi: 10.1016/j.eswa.2017.05.028.

[5] A. Amin et al., "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, vol. 237, pp. 242–254, 2017, doi: 10.1016/j.neucom.2016.12.009.

[6] S. F. Bilal, A. A. Almazroi, S. Bashir, F. H. Khan, and A. A. Almazroi, "An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," PeerJ Comput. Sci., vol. 8, 2022, doi: 10.7717/PEERJ-CS.854.

[7] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," Data Knowl. Eng., vol. 63, no. 2, pp. 503–527, 2007, doi: 10.1016/j.datak.2007.03.016.

[8] F. Yang, "Data-centric AI : Perspectives and Challenges," pp. 945–948.

[9] "Data-Centric AI vs. Model-Centric AI · Introduction to Data-Centric AI." Accessed: Nov. 19, 2023. [Online]. Available: https://dcai.csail.mit.edu/lectures/data-centric-model-centric/

[10] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," J. Artif. Intell. Res., vol. 70, pp. 1373–1411, 2021, doi: 10.1613/JAIR.1.12125.