Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning

Diaa Azzam
School of ITCS
Nile University
Giza, Egypt
di.ayman@nu.edu.eg

Manar Hamed
School of ITCS
Nile University
Giza, Egypt
man.adel@nu.edu.eg

Nora Kasiem
School of ITCS
Nile University
Giza, Egypt
no.abdelhady@nu.edu.eg

Yomna Eid
School of ITCS
Nile University
Giza, Egypt
yeid@nu.edu.eg

Walaa Medhat
School of ITCS
Nile University
Giza, Egypt
wmedhat@nu.edu.eg

***Abstract*—Customer churn poses a formidable challenge within the Telecom industry, as it can result in significant revenue losses. In this research, we conducted an extensive study aimed at developing a viable customer churn prediction method. Our method utilizes the Apriori algorithm's strength to identify the key causes of customer churn. In the pursuit of this goal, we utilized multiple machine learning predictive models. All of which were developed from the insights gleaned from the Apriori algorithm's feature extraction for churning customers. This extensive analysis encompassed a spectrum of machine learning techniques that include Logistic Regression, Naive Bayes, Support Vector Machines, Random Forests, and Decision Trees. Furthermore, we utilized an ensemble learning approach to enhance the predictive accuracy of our models. We also used a voting classifier refined with the best features within our dataset. The voting classifier yielded an accuracy rate of 81.56%, underscoring the effectiveness of our approach in addressing the critical issue of customer churn in the Telecom industry.**

***Index Terms*—Telecommunication Industry, Customer Churn, Classification, Association rules.**

I. Introduction

Customer churn prediction is a crucial study subject with significant business ramifications in numerous sectors. The consequences of customer turnover are severe since it can lead to significant revenue loss, as demonstrated by earlier studies [1], [2]. Given the potential for yearly losses of millions of dollars, client turnover has increased significance in the context of the telecommunications industry.

The identification of clients that are potentially at possibility of churning assumes fundamental relevance inside this complex ecosystem. With this foresight, businesses may proactively engage in client retention measures like personalised service improvements. This tactical move has the potential to reduce client attrition, protecting revenue sources and preserving corporate viability.

In the telecommunications industry, several research on customer attrition prediction have been undertaken. Machine learning approaches such as random forest were employed in one research to predict customer attrition [3]. Another study offered a unique cost-sensitive methodology for predicting client attrition using machine learning [4].

Ensemble methods have also been used to improve the accuracy of customer churn prediction models. The authors in [5] conducted a comparative analysis of state-of-the-art ensemble methods in customer churn prediction. They found that ensemble methods, such as bagging and boosting, can significantly improve the performance of customer churn prediction models. Data preprocessing techniques have also been explored to improve the accuracy of churn prediction models [6].

Several other studies have used various algorithms, such as neural networks and decision trees, to predict customer churn [7], [8]. Additionally, some studies have proposed combining different algorithms to improve churn prediction accuracy [9].

Ensemble learning techniques have taken centre stage as a formidable tool for improving the accuracy of customer churn prediction models. The authors in [5] undertook a comparative analysis to investigate the effectiveness of modern ensemble methods in the context of customer churn prediction. Their thorough analysis revealed how ensemble methods, which include methods like bagging and boosting, can significantly improve the forecasting performance of customer churn models.

Furthermore, as described in [6], the search for greater churn forecast accuracy has sparked investigation into the field of data preparation approaches. The goal of this project is to effectively utilise the synergy between diligent data refinement and predictive model effectiveness.

The research landscape extends further, embracing a myriad of algorithmic approaches. Various studies, as articulated in [7], [8], have rigorously examined the applicability of neural networks, decision trees, and a multitude of other algorithms in the realm of customer churn prediction. Notably, some scholars have advanced the paradigm by advocating for hybrid approaches, where different algorithms coalesce harmoniously, presenting a promising avenue for augmenting churn prediction accuracy, as elucidated in .

The research field is larger and includes a wide range of computational strategies. Numerous research have thoroughly investigated the use of neural networks, decision trees, and a variety of other algorithms in the field of predicting customer turnover, as stated in [7], [8]. As explained in [9], it is noteworthy that certain researchers have enhanced the paradigm by supporting hybrid approaches, where many algorithms come together harmoniously.

One of the essential factors in customer churn prediction is customer value. Therefore, the main contribution of this research is that we have used Apriori Algorithms to extract the most important features then we tested six classic machine learning methods for the prediction task. A comparison of their results is discussed. Moreover, we conducted ensemble learning using the extracted features.

The paper is organized as follows. First, section two provides an overview of current related research. The experiments were then thoroughly explained in section three. The outcomes of the evaluated designs are presented in Section four along with a commentary. The work done was concluded in section five of the paper.

II. Literature Review

In the pursuit of customer churn prediction, numerous innovative techniques, including machine learning, deep learning, and analytical methodologies, have been investigated in the quest to forecast customer attrition [10]. The methodologies utilised in the telecom business are clarified by a thorough ten-year retrospective in [11], which emphasises the need for suitable model selection, optimised data preprocessing, and feature selection to raise prediction accuracy.

These methods have been categorized into four key groups, which are statistical, machine learning, data mining, and hybrid approaches [12]. This taxonomy not only helps navigate the research landscape but also underscores the dynamic synthesis of methodologies. This is in the efforts that are at addressing the complex challenge of customer churn prediction.

A comparative study performed in [13] discussed various techniques used for analyzing customer churn in the field of customer relationship management. The authors compared and evaluated the effectiveness of different techniques such as extreme gradient boosting (XGBoost) and artificial neural networks in predicting customer churn. They analyze the performance of these techniques based on metrics such as accuracy, precision, recall, and F1 score. The study concluded that artificial neural networks outperform other techniques in predicting customer churn.

Another approach was utilized in [14] that used the MLP approach. The approach encompasses a multilayer perceptron neural network (MLP) for predicting customer churn. The MLP is trained on a dataset of historical customer data, and it is able to identify the factors that are most predictive of churn. The MLP was able to predict customer churn with an accuracy of 86%. Back propagating Artificial Neural networks (ANN) were also used in [15]. The proposed neural network-based approach for predicting customer churn. The neural network is trained on 2 real customer datasets, which are IBM Telco and Cell2cell.

The neural network was able to predict customer churn with an accuracy of 86.57%. The Back propagating Neural Network outperformed the XG-Boost-based algorithm. Supervised Machine learning methods form an important approach in customer churn prediction. However, data imbalance, limitations, and various other challenges are faced when using these approaches. Machine learning-based approaches such as decision trees, random forests, and support vector machines were utilized for predicting customer churn in big data platforms in [16]. Their proposed approach was able to predict customer churn with an AUC ranking value of 87.76

In [17] an empirical research study in the telecom industry in China was conducted, proposing a prediction model of customer churn that considers customer value. Similarly, [16] used big data platforms and machine learning to predict customer churn based on customer value in the telecom industry. In [18], [19] described a case study of a telecommunications company, where machine learning algorithms were also utilized to predict customer churn. At the end, the accuracy of the classifiers was evaluated using the test data technique. Then, for producing predictions based, such as logistic regression, decision trees, random forests, and support vector machines, to predict customer churn.

To forecast customer turnover, this study [20] compared the effectiveness of the Decision Tree and Naive Bayes, two traditional classification techniques. Three sources were used to collect the study's data: user behavior, previous financial transactions, and identification information. Finally, the accuracy of the classifiers was evaluated using the test data technique. Subsequently, the most precise classifier was chosen to provide predictions based on customer data when the values of the response variable are unknown. The study's findings demonstrated the efficacy of Decision Tree and Naive Bayes in forecasting client attrition. Naive Bayes was marginally more accurate than Decision Tree.

Ensemble methods combine multiple base learners to improve the overall performance of the model. The authors [5] compare the performance of a variety of ensemble methods on a dataset of customer data from a telecommunications company. The results of the evaluation show that the ensemble methods achieve significantly higher accuracy than the individual base learners. The best-performing ensemble method was the Random Forest with AdaBoost, which achieved an accuracy of 87%. The authors conclude that ensemble methods are a valuable tool for customer churn prediction.

Ensemble learning improves machine learning algorithms by combining multiple diverse models, leading to increased prediction accuracy, enhanced robustness against overfitting, and better handling of complex and non-linear problems. It also helps in addressing imbalanced datasets and reducing the impact of data noise, resulting in improved generalization. Ensemble learning is a versatile and powerful approach that enhances the overall performance and reliability of predictive models in various domains [21], [22].

III. Methodology

Our research approach is based on an ensemble strategy that systematically combines the advantages of many machine learning algorithms to improve prediction accuracy. The Apriori technique is used in the initial stage to select features in a considered manner, thus shaping the foundation of our analysis. The intricacy and predictive power of our model are then further increased by our approach's inclusion of an ensemble learning technique.

A. Dataset & Feature Analysis

Many datasets have been used to test and train churn prediction models. In this study, we use the IBM Telco Customer Churn dataset that is available by the IBM Watson Analytics community [23], [24]. The data encompasses 7,043 client records. The dataset contains features such as age, gender, income, number of dependents, and the services they have subscribed to. The dataset also contains details on the account duration, contractual terms, chosen payment method, use of paperless billing as well as monthly and overall costs, and finally the data set contains an attribute about whether or not the client has left the company.

In our research, feature analyis played a pivotal role to gain more insights on the dataset. Our approach involved two methodologies, which are correlation analysis and mutual information gathering. We were able to quantitatively evaluate the linear relationships between each feature and client attrition through the correlation analysis. We also used Mutual Information Gathering to simultaneously collect linear and non-linear connections. The mutual dependency between characteristics and churn status was assessed using an information-theoretic approach, revealing complex patterns that might be missed by conventional correlation analysis.

B. Data Preprocessing

In the course of our research, we applied a range of techniques designed to uncover concealed insights within a customer churn dataset. This step, known as data preprocessing, plays a pivotal role in analysing the dataset used in our work. The data preprocessing step is primarily performed to refine raw data, ensuring it undergoes a transformation into a well-structured format that lends itself to subsequent analysis with maximum effectiveness. After loading the dataset we perform several preprocessing steps.

1. **Examination of Dataset Characteristics:** In this step, we conducted a thorough review of the dataset's contents. We aimed to gain a deeper understanding of the data by examining the various attributes and their relevance to our research. Regarding the attributes that were not directly relevant to achieving our goal were identified and then eliminated from the dataset. By streamlining the dataset, it becomes more efficient and focused for later analysis.
2. **Numerical Encoding of Dataset Attributes:** Datasets often contain a mix of data types, including numerical and categorical variables. We carried out the numerical encoding on the attributes of the dataset, in order to ensure homogeneity and compatibility for machine learning methods. Categorical data, which typically consists of non-numeric labels or categories, was also transformed into numerical representations. This transformation makes it possible to conduct mathematical operations on all characteristics evenly, which makes it possible to efficiently create and train machine learning models.
3. **Data Splitting into Training and Testing Sets:** To evaluate the performance of our predictive models accurately, we divided the dataset into two subsets , namely, the training set and the testing set. The training set, which comprised 80% of the data, was used to train our machine learning models.The models used this set as a starting point to discover patterns and connections in the data. The testing set, which was unaltered during the model training phase, was made up of the final 20% of the data. Instead, it evaluated how well our models applied to fresh, untested data.

The use of these preprocessing techniques has proven crucial in assuring the best accuracy and performance for our models. By eliminating auxiliary characteristics, encoding categorical data, and stratifying the data into distinct train and test groups, we were able to build highly accurate models with increased generalisation capabilities for unknown data. The thorough preprocessing procedures made a substantial contribution to the overall efficacy and dependability of our research findings.

C. Models Architecture

To predict customer churn we implemented multiple Machine learning models as shown in Fig [1].The Apriori technique, which has gained popularity for its association rule mining capabilities, is used in this study to efficiently identify frequent item sets and noteworthy patterns in our dataset. The robustness of our predictive modelling technique is then improved by utilising six different classifiers, as shown in Figure [1], creating a thorough analytical framework.

- Logistic Regression Classifier: The machine learning statistical model is employed in this work to perform binary classification tasks. The model's primary objective is to categorize clients based on attributes that have been previously mined from the dataset. By utilizing this approach, the study aims to effectively distinguish and assign clients to specific categories, enhancing the understanding and decision-making process in the given context.
- The K-Nearest Neighbour (KNN): A supervised machine learning method used for prediction tasks is the classifier. It works by scanning the training set for the k data points that, according to some distance metric, are the most comparable to the incoming input. The classifier then labels the new data point with the label that appears most frequently among these k neighbors, giving it a flexible and efficient tool for different categorization tasks. It is favored in many real-world applications due to its clarity and interpretability.
- Support Vector Machine (SVM) Classifier: We utilize this classifier for finding the optimal hyperplane to distinguish between two groups of data points effectively. By maximizing the margin between the classes, SVM can efficiently separate the data points and make accurate binary classifications. This approach is widely applied in various

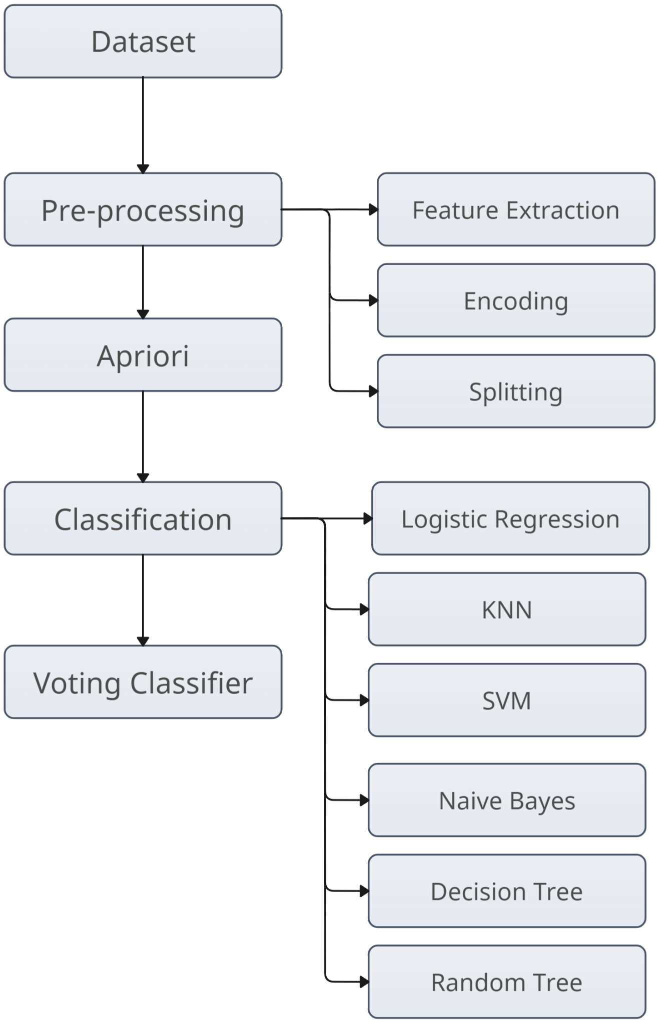

Fig. 1. Proposed system architecture

domains where clear decision boundaries are crucial for successful data classification and pattern recognition.

- Naive Bayes Classifier: Founded on the Bayes theorem, a probabilistic theory, using the erroneous assumption that a data point's properties are independent of one another.
- Decision Tree Classifier: We use this classifier due to its hierarchical structure that provides further insights on the factors affecting customer attrition.
- Random Forest Classifier: We utilized the Random Forest classifier for its ability in capturing non-linear relationships. It creates an ensemble of decision trees from various features, providing accurate predictions and feature importance insights.

All the previously mentioned classifiers were used as the base classifiers for hard voting in the voting classifier. The voting classifier itself is an ensemble learning approach. We utilize it in this study in order to obtain the final prediction by aggregating the results of multiple base classifiers.

IV. Experimental Results & Discussion

To find the most pertinent characteristics in this study, the Apriori method was used with a minimum support of 70% and a minimum threshold of 1. These features were subsequently used to train the classifiers, and the resulting results are shown in Figure [2]. Notably, the voting classifier demonstrated remarkable performance, achieving an accuracy of 81.56% by leveraging the best features extracted through the Apriori algorithm.

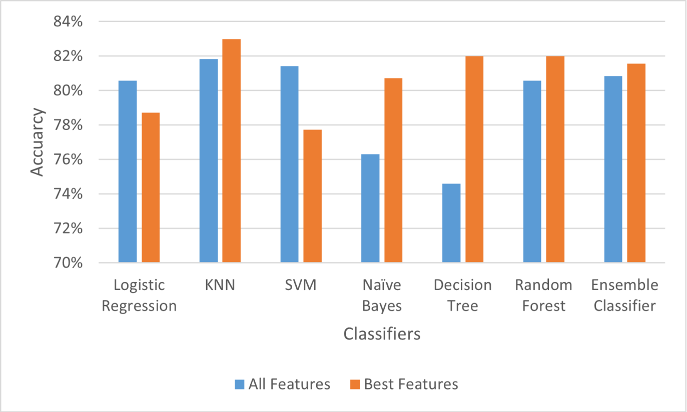

Fig. 2. Results comparison for the proposed Classification Models

In Table [1], the "Features Accuracy" denotes the overall accuracy of our selected models, encompassing the evaluation of various features. Meanwhile, the term "Best Features Accuracy" corresponds to the models' accuracy achieved by employing the Apriori algorithm's selected best features. We also discuss the results of the Voting classifier, denoted in the table as the Ensemble Classifier. In Table [2], we showcase the association rules, confidence, and lift metrics. This is to provide insights into item relationships and patterns. The relationship between service and tenure indicates that people who have been with a phone service provider for a long time are more likely to have a phone service. Table [2] suggests that the best features were service, charges, and tenure.

In this research, we presented a reliable technique that outperforms alternative approaches in terms of scale and reliability. This is performed by the comb the Apriori algorithm with a variety of machine learning algorithms and ensemble learning techniques. Our method combined rule-based association mining with predictive capability while taking advantage of generalization, interpretability, and pattern discovery. In addition to ensuring the adaptation to different data types, the diversity in algorithms used ensures model reliability through ensemble learning. Additionally, Our feature selection step is supported by Apriori's association rule mining, resulting in data-driven improvements.

V. Conclusion

Various methods were investigated to build a model that aims to predict customer churn on the IBM dataset. In this work, we utilized the Apriori algorithm to mine for the

TABLE I
ACHIEVED ACCURACY COMPARISON USING ALL FEATURES AND BEST FEATURES

Classifier	***All Features ACC.***	***Best Features ACC.***
Logistic Regression	80.57%	78.72%
KNN	81.84 %	82.98 %
SVM	81.42 %	77.73%
Naive Bayes	76.31 %	80.71 %
Decision Tree	74.61 %	81.99 %
Random Forest	80.57 %	81.99 %
Ensemble Classifier	80.85 %	81.56 %

TABLE II
CONFIDENCE AND LIFT METRICS COMPARISON

Antecedents	**Consequents**	**Confidence**	**Lift**
Service	Tenure	0.998	1.000147
Service & Charges	Tenure	0.998	1.000147
Service	Tenure & Charges	0.998	1.000147
tenure	Service	0.903	1.000147
Tenure & Charges	Service	0.903	1.000147

best features and then subjecti them to multiple classifiers for performance evaluation. The voting classifier achieved the best result with an accuracy of 81.56%. In conclusion, this study explored diverse approaches to construct a reliable predictive method for customer churn using the IBM dataset. As machine learning and deep learning models continue to progress, our study will eventually incorporate a wider variety of algorithms and approaches in order to better anticipate customer churn. These initiatives will help to advance and hone predictive analytics in the areas of client retention and business intelligence.

REFERENCES

[1] S. J, C. Gangadhar, R. K. Arora, P. Renjith, J. Bamini, and Y. devidas Chincholkar, "E-commerce customer churn prevention using machine learning-based business intelligence strategy," *Measurement: Sensors*, vol. 27, p. 100728, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917423000648

[2] A. T. Institute, "Churn meaning," 2023. [Online]. Available: https://ati.ac/churn-meaning/

[3] I. Ullah, B. Raza, A. Malik, M. Imran, S. Islam, and S. W. Kim, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. PP, pp. 1–1, 05 2019.

[4] A. Correa Bahnsen, D. Aouada, and B. Ottersten, "A novel cost-sensitive framework for customer churn predictive modeling," *Decision Analytics*, vol. 5, pp. 1–15, 06 2015.

[5] M. Bogaert and L. Delaere, "Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art," *Mathematics*, vol. 11, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/5/1137

[6] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 03 2021.

[7] A. Sharma and D. Panigrahi, "A neural network based approach for predicting customer churn in cellular network services," *International Journal of Computer Applications*, vol. 27, 09 2013.

[8] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a customer churn combination prediction model based on decision tree and neural network," in *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 04 2020, pp. 129–132.

[9] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 39, no. 1, p. 1414–1425, jan 2012. [Online]. Available: https://doi.org/10.1016/j.eswa.2011.08.024

[10] S. Min, X. Zhang, N. Kim, and R. Srivastava, "Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets," *Journal of Marketing Research*, vol. 53, 02 2016.

[11] N. Hashmi, N. A. Butt, and D. Iqbal, "Customer churn prediction in telecommunication a decade review and classification," *IJCSI*, vol. 10, pp. 271–282, 09 2013.

[12] S. De, P. P, and J. Paulose, "Effective ml techniques to predict customer churn," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 895–902.

[13] O. Celik and U. OSMANOGLU, "Comparing to techniques used in customer churn analysis," *Journal of Multidisciplinary Developments*, vol. 4, no. 1, 2019.

[14] O. Adwan, H. Faris, K. Jaradat, O. Harfoushi, and N. Ghatasheh, "Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis," *Life Science Journal*, vol. 11, pp. 75–81, 01 2014.

[15] W. Fujo, S. Subramanian, and M. A. Khder, "Customer churn prediction in telecommunication industry using deep learning," *Information Sciences Letters*, vol. 11, pp. –, 2022. [Online]. Available: https://digitalcommons.aaru.edu.jo/isl/vol11/iss1/24

[16] A. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, 03 2019.

[17] A. Farouk, M. Zhao, Q. Zeng, M. Chang, Q. Tong, and J. Su, "A prediction model of customer churn considering customer value: An empirical research of telecom industry in china," *Discrete Dynamics in Nature and Society*, vol. 2021, p. 7160527, 2021. [Online]. Available: https://doi.org/10.1155/2021/7160527

[18] P. Lalwani, M. Mishra, J. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, pp. 1–24, 02 2022.

[19] B. Senthilnayaki, S. M, and N. D, "Customer churn prediction," *IARJSET*, vol. 8, pp. 527–531, 2021.

[20] C. CIMPOERU and A. Andreescu, "Predicting customers churn in a relational database," *Informatica Economica*, vol. 18, pp. 5–16, 09 2014.

[21] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157823000228

[22] T. Akano and C. James, "An assessment of ensemble learning approaches and single-based machine learning algorithms for the characterization of undersaturated oil viscosity," *Beni-Suef University Journal of Basic and Applied Sciences*, vol. 11, 12 2022.

[23] yeanzc, "Telco customer churn," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset

[24] BlastChar, "Telco customer churn," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn