UNIVERSITÀ DEGLI STUDI ROMA TRE

Master's Degree in Computer Science Engineering

Data-Centric AI to Churn Prediction with Synthetic Data

Alessandro Pesare
553700

A.Y. 2023/2024

Contents

Contents **ii**

1 Introduction **1**

2 Tabular data synthesis algorithms **3**

2.1 Syntetic Data 3

2.2 Data synthesis algorithms 4

3 A Comparative study **7**

3.1 Experimental settings 7

3.2 Results 9

Conclusions **10**

Bibliography **11**

Chapter 1

Introduction

Churn prediction, which aims to identify customers that are going to leave in advance so the service provider can apply their campaign strategy to retain these customers, is of significant importance for subscription-based industries, such as insurance, telecommunication, and others. This is a critical operation for many businesses because acquiring new clients often costs more than retaining existing ones. Therefore, being able to detect customer churn early and take marketing actions based on artificial intelligence (AI) systems is essential for businesses. In general, algorithms and data are two integral components of any AI system. The research of AI systems can be classified into two groups: model-centric AI and data-centric AI[8]. While model-centric AI is developed to improve the performance of specific models, data-centric AI aims to improve the quality of the data for downstream machine learning tasks. On the other hand, the data-centric approach is the opposite, where the task is to improve existing data or integrate new data, then train and evaluate the ML algorithms. Even though we have data created consistently on a daily basis, there are still practical constraints when using real data. These limitations range from the need for big data, the difficulty of obtaining specific data, and privacy concerns when the data is sensitive. Churn data is a typical example, which often involves all three data limitations. First, churn data is a valuable asset that has ownership and business boundaries. This business boundary leads to a data availability issue where data with personal information is generally limited to certain groups due to business concerns or regulatory requirements. Second, in real-world churn datasets, the churn variable class is usually imbalanced, such that the ratio of the churner class is often much lower compared to the non-churner class. Therefore, the skewed class proportions require more training data from the churner class to have better model performance. Third, even with a substantial amount of churn data collection, there is a significant business cost to labeling the churner, given customer churn poses a significant revenue loss to the business. Given the business significance of the churn prediction, it is a widely researched topic, where numerous model-centric AI systems have been developed.

The main focus of this work is to investigate whether we can improve churn prediction by substituting, balancing, and augmenting real data with data synthesis.

It is worth noting that the experiments described in this work are extensively detailed in the paper *"Data-Centric AI to Improve Churn Prediction with Synthetic Data"* [6] , which I followed to delve deeper into the topic of synthetic data within the context of data-centric AI.

Chapter 2

Tabular data synthesis algorithms

2.1 Syntetic Data

In churn prediction[5], Machine Learning models are trained using past data to make predictions about future events. This involves two main components: the model itself, consisting of the algorithm and its hyperparameters, and the dataset used for training. While the model aspect is largely developed and widely available, the focus now shifts to the data, which is where data-centric AI comes into play. Synthetic data refers to artificially generated data that mimics the statistical properties of real-world data. It is created through data synthesis techniques, where models are trained to understand and replicate the patterns and characteristics of authentic datasets. Synthetic data[3] serves as a valuable resource in situations where obtaining or using real data is impractical or restricted due to various reasons such as privacy concerns, data scarcity, or confidentiality issues. In churn prediction and other machine learning tasks, synthetic data offers a promising solution to address challenges related to data quantity, quality, and availability. By simulating realistic data samples, that can effectively supplement existing datasets, enhance their utility and augment data when additional samples are needed to expand the size of the dataset for better model performance. Synthetic data mirrors real datasets by learning statistical properties from the original data through a process called data synthesis. This involves training a model to understand the distribution of real data and using it to generate synthetic data. Consequently, synthetic data can be tailored to address specific needs, such as filling in missing values or adjusting extreme values in tabular data. Overall, data synthesis offers the potential to enhance datasets by providing additional data, balancing class distributions, and preserving confidentiality when dealing with sensitive information.

2.2 Data synthesis algorithms

Creating tabular synthetic data can be approached in various ways. One method involves using perturbation-based models, which adjust the values in existing tables through linear or non-linear transformations. This method is relatively straightforward to implement, especially for small datasets with continuous features. However, because each synthetic row is generated based on a corresponding row in the real data, this approach doesn't increase the dataset's size or provide robust privacy protection. Moreover, more research is required for this method to handle categorical features and dependencies between features. Another approach involves manually specifying transformations based on domain expert knowledge. While this method allows for more tailored adjustments, it can be costly and complex to develop. Both perturbation-based and manual specification methods often require significant user interaction and expertise, which can limit their generalization across different datasets. To address these challenges, it's advantageous to learn the synthesis rules directly from the training data, moving closer to an end-to-end learning approach. By doing so, we can automate the process and adapt to diverse datasets more effectively. Three popular methods that align with this principle are:

- **Distance-based algorithms**: These algorithms utilize distance measures like Euclidean distance, which are common in many machine learning (ML) algorithms. The Synthetic Minority Oversampling Technique (SMOTE) is one such algorithm originally developed to increase the representation of minority classes by generating synthetic data points based on the k-Nearest Neighbor (k-NN) algorithm. However, SMOTE has encountered two main limitations. Firstly, it tends to introduce noise into the dataset by creating uninformative or noisy samples. Secondly, SMOTE is primarily designed for continuous variables, given that calculating the distance between two categories can be complex. To address this, the Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTENC) was introduced, in which the median of all continuous variable standard deviations is added to the distance metric if the nominal features differ between a sample and its potential nearest neighbors. While SMOTE and its variants are effective in generating synthetic samples, they have been criticized for their reliance on local information rather than capturing the overall joint distribution of the data.

- **Statistical joint-distribution-based algorithms**: These algorithms offer another approach to synthesizing tabular data by leveraging knowledge of the joint distribution present in the real dataset. One popular model in this category is the Copula, which is adept at explaining dependencies between multiple variables without imposing restrictions on their marginal distributions. However, it's worth noting that Copula models are predominantly studied for continuous variables. Synthpop is another popular method from this group, designed primarily to protect confidentiality by constructing joint probability distributions through a one-on-one regression model. When dealing with only continuous variables, Synthpop utilizes sequential conditional distribution. In cases where both continuous and categorical variables are present, it employs the classification and regression tree (CART) technique. While algorithms in this group offer simplicity and flexibility, they are primarily tailored for continuous data, resulting in degraded performance when applied to heterogeneous tabular datasets.

- **Deep generative models**: This techniques represent another strategy for generating synthetic data. Generative adversarial network (GAN)[4] stands out as one of the most versatile neural network architectures since its proposal in 2014. GANs have found successful applications across various domains, including computer vision and natural language processing. Another interesting algorithm in this category is the Variational Autoencoder (VAE), which focuses on building a latent space to encourage efficient data representation. Specifically for tabular data, Conditional Tabular GAN (CTGAN)[7] and Triplet VAE (TVAE)[2] have emerged as popular deep generative models. These models offer flexibility and strong performance in representing tabular data without the need for prior distribution assumptions. Additionally, they provide the capability to generate large amounts of synthetic data once trained, making them appealing options for data synthesis tasks.

Chapter 3

A Comparative study

In this chapter, following a review of some established strategies for synthetic data generation documented in literature, we present the findings from experiments conducted in the paper "Data-Centric AI to Improve Churn Prediction with Synthetic Data"[6]. The primary aim of this study is to analyze and benchmark the most effective data-centric resampling methods for churn prediction

3.1 Experimental settings

The overall experimental framework is showed in Fig.3.1, the goal is investigate the potential of leveraging synthetic data to enhance churn prediction performance compared to a baseline. Baseline approach involves training machine learning (ML) models on real training data and evaluating them on a test set. Experiments are divided into three main parts, each addressing a different data strategy:

- Train ML models on synthetic data and assess their performance on the real test data. Can synthetic data serve as a substitute for real data?
- Train ML models on balanced real training data, incorporating minority classes from synthetic data, and evaluate them on the real test data. Can synthetic data be used for data balancing?
- Train ML models on augmented real training data, which includes synthetic

data, and evaluate them on the real test data. Can synthetic data aid in data augmentation?

In this study, authors evaluate seven well-established tabular data synthesis models, including SMOTE[1], SMOTENC, SMOTE-Tomek, Copula, Synthpop, CTGAN[7], and TVAE[2], across three data strategies: data synthesis, data balancing, and data augmentation comparing these models with a baseline approach over four datasets. Evaluation is based on test sets comprising 20% of the original dataset, completely separated from the training set (80%). Synthetic data is generated solely from the training set to prevent information leakage. The area under the precision-recall curve (AUCPR) serves as the primary evaluation metric due to the imbalance in most churn datasets. To ensure robustness, experiments are conducted with the same settings but repeated five times with different random seeds, and the average AUCPR is reported. For simplicity, default hyperparameters are used for both data generation and classification algorithms.

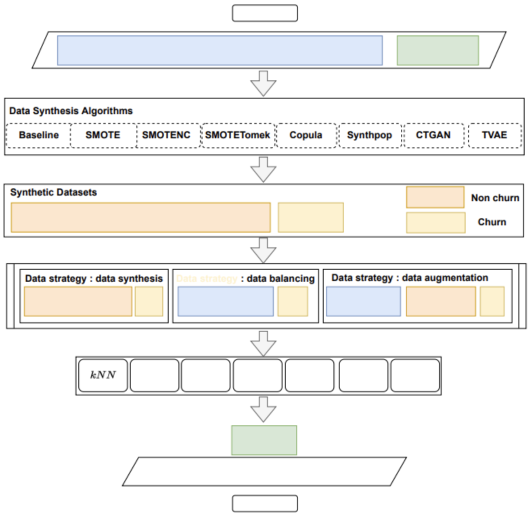
Data Synthesis Algorithms
Baseline
SMOTE
SMOTENC
SMOTETomek
Copula
Synthpop
CTGAN
TVAE
Synthetic Datasets
Non churn
Churn
Data strategy : data synthesis
Data strategy : data balancing
Data strategy : data augmentation
kNN

Figure 3.1: Experimental framework for comparing data synthesis models and determining the effects of data strategies

3.2 Results

To examine various data generation algorithms and evaluate their impact on churn prediction, a competitive analysis was conducted. The study compared the baseline approach against three distinct data strategies: data synthesis, data balancing, and data augmentation. The findings are depicted in Fig.3.2, where the average Area Under the Curve of the Precision-Recall curve (AUCPR) was analyzed using a spider chart. This chart illustrates different data generation algorithms, including SMOTE, SMOTENC, SMOTETomek, Copula, Synthpop, CTGAN, and TVAE, positioned along the outer ring. The performance metric, average AUCPR, is represented on the axis, ranging from 0.5 to 0.7, with higher values indicating superior performance. Each data strategy is differentiated by a distinct color. Overall, both data balancing and data augmentation resulted in enhancements to the baseline classification performance (0.6671), underscoring the effectiveness of data-centric AI in churn prediction. Among the data synthesis models, TVAE exhibited the most favorable outcomes with both data balancing (0.6733) and data augmentation (0.6749), positioned farthest from the chart's center. SMOTENC, SMOTETomek, and SMOTE with data balancing also showcased improvements over the baseline (0.6704, 0.6675, and 0.6675, respectively).

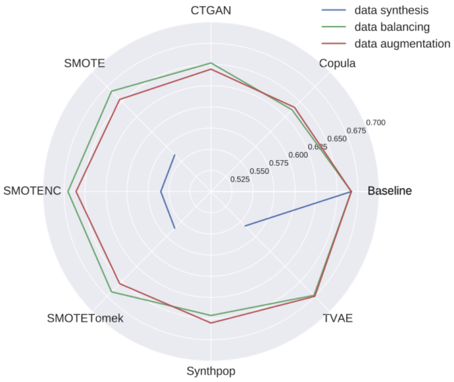

Figure 3.2: Experimental framework for comparing data synthesis models and determining the effects of data strategies

Conclusions

The utilization of synthetic data in the realm of data-centric AI presents a compelling avenue for exploration and innovation. On one hand, the iterative approach to improving data quality, typical of data-centric AI approaches, is crucial for enhancing accuracy through techniques such as data cleaning or augmentation. On the other hand, it's equally intriguing to consider the role of synthetic data within this context. The utilization of synthetic data can bring significant value as it allows us to maintain a focus on data while leveraging techniques inherited from machine learning or deep learning to enhance their quality. Delving into the realm of enhancing synthetic data holds promise for several reasons. Firstly, it offers a unique opportunity to challenge and refine AI models in scenarios where real-world data is scarce, biased, or sensitive. By iteratively improving the quality and diversity of synthetic data, we can create more robust AI systems capable of navigating complex and varied real-world environments. Secondly, leveraging synthetic data opens doors to experimentation and scenario testing that might be unfeasible or unethical with real data. This allows researchers and practitioners to explore the bounds of AI applications without risking privacy breaches or legal constraints. By generating synthetic data that retains the statistical properties of real data while ensuring individual privacy, we can unlock the potential of sensitive datasets for research and innovation while safeguarding personal information. In conclusion, the pursuit of enhancing synthetic data within the data-centric AI landscape offers many great opportunities for progress and to tackle the complexities of our rapidly evolving world.

Bibliography

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. 2011.

[2] H. Ishfaq, A. Hoogi, and D. Rubin. Tvae: Triplet-based variational autoencoder using metric learning, 2018.

[3] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, and W. Wei. Machine learning for synthetic data generation: A review, 2023.

[4] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. Data synthesis based on generative adversarial networks. 2018.

[5] B. Prabadevi, R. Shalini, and B. Kavitha. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4:145–154, 2023.

[6] A. X. Wang, S. S. Chukova, and B. P. Nguyen. Data-centric ai to improve churn prediction with synthetic data. In *2023 3rd International Conference on Computer, Control and Robotics (ICCCR)*. IEEE, Mar. 2023.

[7] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan, 2019.

[8] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, and X. Hu. *Data-centric AI: Perspectives and Challenges*, page 945–948. Society for Industrial and Applied Mathematics, Jan. 2023.