Synthetic Data Generation: Generating High-Utility Synthetic Parcel Data

CS4991 Capstone Report, 2021

Janessa Jiang
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
jj8bp@virginia.edu

ABSTRACT

In an age of rapid technological advancement, significant progress continues to be made in the field of artificial intelligence. Accompanying this progress is the application of artificial intelligence in various fields and industries. In particular, machine learning, a subcategory of artificial intelligence, has been used to create synthetic data. While interning on a software development team at Amazon Web Services (AWS), I was assigned a project that involved the utilization of AWS technologies to generate synthetic customer telemetry parcel data to increase test data accessibility and address customer privacy concerns.

Amazon AutoGluon, an open-source machine learning library, was used within a notebook instance to build and train data models. These models were used to predict values and generate high-utility synthetic parcels. The resulting synthetic parcel data closely followed the trends and relationships between attributes that were present in real parcel data. Evaluation of the data utility involved statistical calculations, assessment of data distribution, and testing. The data models met the project objectives by allowing users to generate anonymized parcel data at any volume. This project opened doors to the possibility of applying machine learning techniques to existing AWS services to improve testing and other parts of the development pipeline.

1 Introduction

As the fields of artificial intelligence and machine learning continue to advance, the applications of these technologies are becoming more widely-accepted and prevalent in society. Synthetic data, data that is artificially generated, has use cases in many fields, such as healthcare and transportation. This data proves useful for detecting anomalies, providing unlimited testing data, and other use cases. Used in place of real data, synthetic data also addresses privacy concerns by providing anonymity.

As an internal team that seeks to properly calculate the amount of usage of AWS offerings by customers to determine billing, it is important to be both accurate and efficient. As the technologies used on this team evolve and improve over time, it is necessary to continually test to ensure quality and precision in calculations. The usage of synthetic data allows for unlimited testing data and full anonymity which makes continuous testing feasible.

2 Background

Customer billing data and usage statistics are bundled into packages called parcels. These parcels serve as input for testing the billing calculation services used on this team. As shown in Figure 1, real parcel data is currently anonymized to protect customer personal data. However, anonymized data has the risk of re-identification. Additionally, real data is limited and not readily available. Through the usage of synthetic data, these concerns may be mitigated. The number of parcels needed for testing can be adjusted based on the testing needs, and re-identification is no longer a risk.

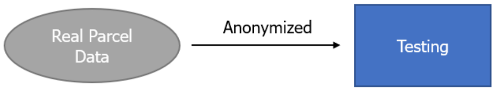

Figure 1: Current Parcel Testing Pipeline (Jiang, 2021)

Parcel data consisted of identifiers, service usage data, location of data centers, data plans, and other information. Pricing calculations vary from service to service and also depends on the pricing plan chosen by the customer. Three offered payment plans are shown in Figure 2. For example, a customer may choose a pay-as-you-go plan or save money by committing to use a specific amount of an AWS service over a one- or three-year period. All of these factors and complicated relationships between variables must be properly represented in synthetic parcel data.

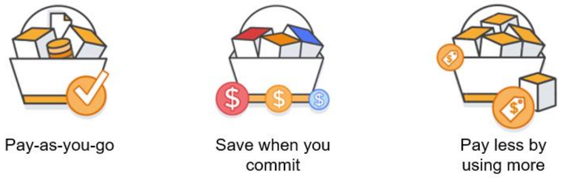

Figure 2: Various AWS Billing Plans (AWS, 2021)

3 Review of Research

Synthetic data continues to gain traction and widespread acceptance in a variety of fields, including healthcare and transportation. Concerns that hindered progress, such as privacy worries and having limited data, can be mitigated through the usage of synthetic data. Hittmeir et al. [2019] delves into the usage of synthetic data as a main disclosure control measure in their research due to the increasing need to keep personal data secure [1]. Synthetic data proves more secure than traditional anonymization methods because it eliminates the risk of re-identification. In cases where real data is not readily accessible, synthetic data serves as a valuable and abundant replacement. Lack of access to testing data may slow down the research and development process. Using synthetic data will address this concern.

As the usage and application of synthetic data becomes more prevalent, there is growing skepticism over the usage of synthetic data in replacement of real data. To prove that synthetic data generated by machine learning techniques is usable, its utility must be evaluated with both general and specific measures [2]. Appropriate measures of utility may vary across fields due to different use cases for synthetic data. When assessing identification disclosure risks, steps may be taken to improve the model if privacy is still a concern. Reiter and Mitra [2009] examine methods to eliminate risk of re-identification and increase anonymity for partially synthetic data in their research [3]. Conducting both utility evaluations and privacy assessments will help synthetic data gain acceptance.

4 Project Design

The process of model training and synthetic data generation, as well as relevant challenges, are represented in subsections 4.1, 4.2, and 4.3 below.

4.1 Data Model Generation Process

Anonymized real data was used to train models within a notebook instance. This parcel data was pulled from an AWS Simple Storage Service (AWS S3) bucket and was in the form of nested JSON. Using the pandas DataFrame library, the parcel data was massaged and transformed to remove outliers and fill in missing values. After mapping the parcels to a dataframe, the dataframe served as input to build and train data models. Amazon AutoGluon, a machine learning library, automates machine learning tasks. In this case, it was used for multi-label tabular prediction. The relationships between each of the table attributes were examined in order to train useful models. The library runs as many algorithms as it can during the period of training time and chooses the most appropriate algorithm for that attribute model. Figure 3 outlines this process for generating synthetic parcel data.

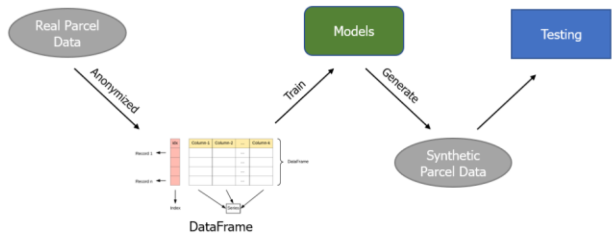

Figure 3: Proposed Parcel Testing Pipeline (Jiang, 2021)

4.2 Synthetic Parcel Data Generation

The models trained using Amazon AutoGluon were used to generate synthetic parcel data. This fitting and predicting process is shown in Figure 4. After generating this data, it was reverted back to the formatting of real parcels before being uploaded to a destination AWS S3 bucket. The pandas DataFrame library was utilized to nest the JSON data. These synthetic parcels were evaluated for utility and consistency. The distributions and statistics of real parcel data was compared to that of the synthetic parcel data.

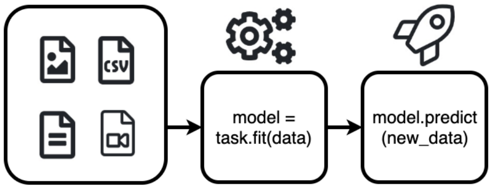

Figure 4: Process of Training Models and Predicting Data (Erikson, 2019)

4.3 Challenges

There was a significant number of attributes per parcel. The process to map the parcel data to a dataframe and flatten the nested JSON was time-consuming and inefficient. The resulting table was very large and memory-intensive. Running the algorithms to create the models took over an hour. Since time-efficiency was a priority of the team, setting aside an hour of time for training models was a roadblock in this process.

Due to the complexity of the relationships between the variables, it was difficult for the machine learning libraries to understand and replicate these trends and distributions. Using the utility measure of comparing synthesized data to real data, the level of utility was low and not ideal for testing purposes.

5 Results

Usage of synthetic parcel data helped to resolve the project objectives set by my team. Specifically, the risk of re-identification of anonymized real parcels was mitigated, and the number of synthetic parcels generated is unlimited. Amazon prides itself on its culture of trust, so maintaining trust with their customers is of utmost importance. As shown in Figure 5, privacy assurance is a necessary part of data synthesis. On this front, using synthetic data for testing purposes in place of real data is a step in the right direction.

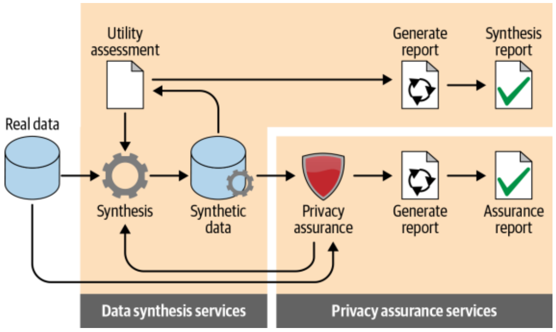

Figure 5: Utility and Privacy Evaluation (O'Reilly, 2020)

However, the nature of multi-label prediction causes data utility to generally be low. Coupled with the complex relationships between the attributes in parcel data, the level of utility for generated synthetic data parcels was lower than anticipated. One method of increasing accuracy is to train models for a longer period of time. However, time-efficiency was a priority for my team. Therefore, it was necessary to find a balance between training time and accuracy. After a certain period of training time, the accuracy of generated data plateaus. This balance would be right before the plateau.

6 Conclusion

As one of the many applications of artificial intelligence, synthetic data has the potential to mitigate concerns in research and development in various fields and industries. As Internet use becomes even more prevalent, it is even more pertinent to protect personal information. Breaking barriers to research and development, such as lack of data, will also be beneficial for the advancement of technology and society. The usage of synthetic data will address both privacy and data scarcity issues. In this case, replacing real parcel data with synthetic testing data for testing purposes is promising. Although results of the utility assessment of synthetic data generated from multi-label tabular models were not ideal, steps can be taken to increase utility. As progress continues to be made in the field of machine learning, tabular prediction will become more accurate and effective for data with many attributes. Using synthetic data in this context proved worthwhile, and opens up doors for the application of this technology in other parts of the development pipeline.

7 Future Work

To emphasize the team's priority of time-efficiency, it will be helpful to conduct further research on maximizing data utility while minimizing time spent on training models. Time-consuming algorithms can be excluded when determining which model best fits each attribute. It would also be helpful to determine the minimum utility necessary for testing purposes in order to precisely calculate the maximum time needed to be spent on model training.

Other automated machine learning (AutoML) libraries can be used to train models and predict values for synthetic data. Future work may involve the comparison of Amazon AutoGluon to other libraries. Since Amazon AutoGluon is a relatively new open-source toolkit, new features continue to be added. As improvements are made, the utility of synthetic data will likely increase, and training time will likely decrease.

REFERENCES

[1] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. 2019. Utility and privacy assessments of synthetic data for regression tasks. *2019 IEEE International Conference on Big Data (Big Data)* (December 2019), 5763–5772. DOI:http://dx.doi.org/10.1109/bigdata47090.2019.9005476

[2] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181, 3 (2018), 663–688. DOI:http://dx.doi.org/10.1111/rssa.12358

[3] Jerome P. Reiter and Robin Mitra. 2009. Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* 1, 1 (2009), 99–110. DOI:http://dx.doi.org/10.29012/jpc.v1i1.567