Privacy First Path Analysis Using Clickstream Data

Krishna Chaitanya Gadepally [1], Sambandh Bhusan Dhal [2,*], Stavros Kalafatis [3] and Kevin Nowka [4]

[1] Department of Electrical and Computer Engineering, Texas A&M University; kcgadepally@tamu.edu
[2] Department of Electrical and Computer Engineering, Texas A&M University; sambandh@tamu.edu
[3] Department of Electrical and Computer Engineering, Texas A&M University; skalafatis-tamu@tamu.edu
[4] Department of Electrical and Computer Engineering, Texas A&M University; kjnowka@tamu.edu
* Correspondence: sambandh@tamu.edu; Tel.: +1 979-739-8326

Abstract: In today's digital economy, data-based decisions have become very important to meet the ever-growing needs of customer engagement, retention, and satisfaction. Clickstream data is one such data that is being used to better understand, predict and engage with customers. Unfortunately, clickstream data for understanding customers has raised privacy and security concerns with many internet providers selling data for monetary benefits. This paper showcases a methodology that is developed based on experiential learning and using the latest cryptographic methods including differential privacy and graph analytics for predicting customer lifetime value (CLV) using clickstream data. Results obtained show that a user's engagement can be predicted within a relatively acceptable range after preserving privacy.

Keywords: Clickstream; RFM; Privacy

1. Introduction

Clickstream analysis is essential in today's world of data driven decisions for enterprises. There are many kinds of clickstream analytics from email marketing, website trends to user churn. Clickstream analysis consists of two parts: clickstream data collection and clickstream analytics. Clickstream analysis is most successful when combined with other, more traditional market research, assessment, data sources, and methods. The main purposes of clickstream data analytics are: operational analytics, ecommerce analytics, marketing research and analytics and customer personalization analytics.

Collecting, analyzing, and reporting aggregate data on which pages a website user visits and in what sequence, for example. In this case, if a large number of users leave a site after landing on a page with insufficient information, the organization may need to improve the page with more useful content.

Because clickstream analysis typically necessitates the collecting of vast amounts of data, many businesses rely on big data technology or third-party providers to analyze the data and create reports on specific areas of interest. More recently, there is a growing trend to use clickstream analysis along with consumer data to provide opportunities for predicting consumer behavior, targeting customers experiences, and providing customized services. In conjunction with clickstream analysis, modern machine learning practices are being introduced like recency, frequency, monetary modeling, churn prediction, classification and many more methods that involve gathering more and more consumer data.

Due to these opportunities and growth in collecting consumer data, organizations are facing many privacy and security challenges. According to the Privacy Rights Clearinghouse, since 2005, 9015 data-breach instances have been made public, affecting billions of sensitive personal details. The scope and magnitude of data breaches are concerning. For example, the 2018 Marriott data-breach event revealed sensitive personal information such as passport numbers, credit and debit card details, and Starwood Preferred Guest

 doi:10.20944/preprints202304.0904.v1

cards, affecting millions of customers. Consumers have raised severe worries about how companies manage their data and preserve their privacy. According to an online study done by IBM in 2018, 78% of US customers claimed that a company's capacity to keep consumer data private is "extremely important," yet just 20% said they "completely trust" the companies with whom they engage to keep their private data safe. Another survey conducted by Consumer Reports found that in the aftermath of Facebook's Cambridge Analytica scandal in 2018, in which the British consulting firm deceptively acquired and used the data of millions of Facebook users, 70% of Facebook users changed their behavior, taking more precautions with their posts, revising privacy settings, and turning off location tracking. These instances demonstrate that, as consumers worry about privacy, they are skeptical of corporations that use their personal information. As a result, corporations are today experiencing a crisis of customer trust and confidence.

Governments are likewise worried about the sufficiency of data security and consumer privacy protection adopted by businesses. Consequently, many new laws like PIPEDA, Indian Privacy laws (Personal Data Protection Bill) and CCPA have been added making it even harder for organizations to work with this data. While giving customers additional rights and protection, such stringent restrictions would inevitably limit corporations' capacity to personalize their marketing operations and services to each consumer. As a result, it is critical to discover solutions that might mitigate the potentially negative side effects of rigorous privacy legislation while maintaining data security.

In this paper, it is shown how organizations can achieve RFM modeling by enabling privacy-preserving storage and access mechanisms to access consumer data. In particular, it is demonstrated how organizations can avoid storing consumer data by applying Federated Learning (FL), clustering and Differential privacy methods while still being able to reach their business goals. The industry is moving towards FL Algorithms on user devices; however, this is still not widely available, and would be available only starting 2023 onwards in a few years [1][2]. The FL technique offers a particular benefit over other privacy-protection strategies. Even if the datasets are mostly anonymized, they can nonetheless jeopardize customer privacy [1]. It demonstrates FL capabilities along with local differential privacy using clustering techniques for predicting customer retention and churn prediction. Our work in this paper is focused on demonstrating the feasibility of applying these methods in data centers along with big data technologies for storing and accessing data.

To show the applicability of the suggested technique in a broad marketing scenario, a consumer browsing a particular dataset from an online store is utilized as an example, with the goal of predicting a consumer's behavior [3]. To establish the FL algorithm, the learning requires the understanding of sensitive attributes, business context and usage patterns on data collectively. It is proved that the suggested approach's prediction accuracy is equivalent to that of the centralized learning method. As a result, this technique enables organizations to target customers with great precision while maintaining personal data protection.

The rest of this paper is structured as follows. In the following section, we briefly discuss the related literature. Section 2 gives an overview of our privacy guarantee objectives. In Section 3, a brief overview of our methodology for storing and accessing data has been discussed. In section 4, we apply our algorithm to a practical marketing problem, training a model to predict each consumer's behavior using an online retailer's clickstream data. Section 5 concludes on the future and benefits of our methodology.

2. Background Literature and Methods

For customer privacy, this article depends on literature. Hann et al. [4] investigate the trade-off that customers confront when deciding whether or not to provide personal information. Hann et al. [4] estimate that the value of personal information protection to people is between $30.49 and $44.62. They discover that perks such as monetary rewards

 doi:10.20944/preprints202304.0904.v1

and future convenience have a considerable impact on users' choices for websites with different privacy rules. Tucker [5] demonstrates that giving consumers greater control over their personal information boosts the efficacy of behavioral targeting. As a result of the privacy rule, Goldfarb and Tucker [6] show that display advertising becomes far less successful (65% less effective on average) in terms of declared purchase intent. Johnson [7] discovers that limited targeting owing to tougher privacy regulations reduces advertiser surplus and, as a result, publishers' profits. Recently, Rafieian and Yoganarasimhan [8] used machine learning techniques to quantify the value of targeting information, specifically the relative importance of contextual information (based on the content of the website and thus preserving privacy) versus behavioral information (based on user-tracking and thus jeopardizing privacy). This work also draws on the literature (e.g., Moe and Fader [9], Montgomery et al. [10], Park and Fader [11], Hui et al. [12]) to analyze customers' decision-making along the buying funnel using path-tracking and click-stream data. In recent years, researchers have focused on privacy-preserving learning. Shokri and Shmatikov [13] present an approach for collaborative learning based on Differential Privacy (DP), in which each person asynchronously trains a neural network locally and selectively communicates just a subset of parameters with other parties. A more comprehensive review and debate on big data and consumer privacy has been provided by Jin [14].

Tracking and keeping journey and click-stream data raises further privacy concerns. Even after anonymization, empirical patterns encoded in data might expose a significant amount of personal information (Valentino-DeVries et al. [15]). This study illustrates that path-tracking and click-stream data may be analyzed without risking users' privacy.

The next sections in this paper introduce the background research for our approach, followed by methodology, results and conclusion. The steps to guarantee privacy and security used are as follows: Discover - Sensitive data discovery, Transform - Differential Privacy and Aggregate - Differential Privacy.

3.1. Privacy and Security issues with Stream Data Processing

- Streaming data processing refers to data that is generated continuously, usually in high volumes and at high velocity.
- Tracking frequent occurrences or situations where events need to be instantly identified and responded to benefit from stream processing. Typically, streaming data consists of log events that capture events as they occur, such as a user clicking on a link in a web page or a sensor reporting the current temperature. Data in its raw form is typically unstructured, semi-structured, or in JSON format. Because of the absence of schema and the volume of data, streaming data is impossible to query using SQL-based analytic tools in its raw form; instead, it must be processed, parsed, and formatted before any significant analysis can be performed. An example architecture of stream processing is as in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

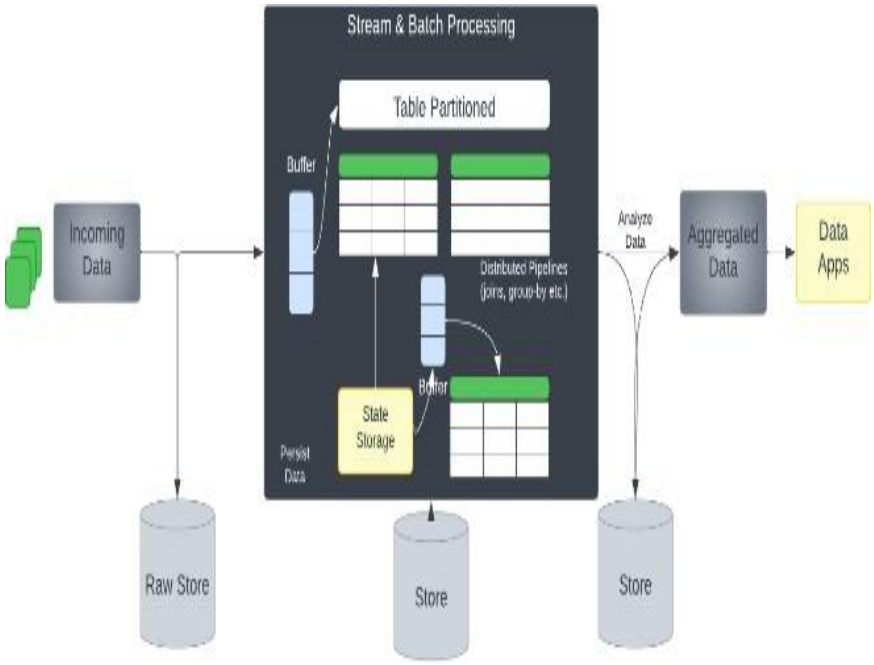

Figure 1. An example of architecture of stream processing

The data storage in the above (Figure 1) architecture is a major concern for privacy and security. Current data ecosystems for streaming do not make this easy to do the same.

3.2. Methods

"Differential privacy" describes a promise, made by a data holder, or curator, to a data subject [16]. The phrase "differential privacy" refers to factors ("epsilon and delta") that quantify "privacy loss"—the increased risk to a person because of data consumption.

Some of the methods in differential privacy methods are: Randomization, Laplace Method and Exponential Mechanism.

3.3. Data Security and Privacy Requirements

No matter whether it is cloud computing or third party computing, the end user's data storage and control mechanisms are separated. This could lead to data leakage, loss, and illegal data operations. Therefore, the privacy protection and security of data are still fundamental problems of streaming data processing. The main challenges are below:

- Confidentiality
- Integrity
- Availability
- Authentication and access control
- Privacy requirements

The focus of our research/study is availability, confidentiality and privacy.

3.4. Data Security and Privacy Challenges

Data streaming utilizes many recent technologies to compute the data as fast as possible or in proximity. This introduces many challenges in data privacy and security like privacy leakage, data tampering and many more. Due to these reasons data security and privacy-preserving have become the basic requirements to protect end users in their business, economics, and daily life.

 doi:10.20944/preprints202304.0904.v1

3.5. Targeted Privacy Guarantees

This methodology is only applied for near real-time data. Near real-time data implies action within one hour. Sensitive data analysis is outside the scope of this paper.

In this paper, the privacy guarantees are designed based on the following need-to-know principles:

- Event Level Privacy: Events should hide the user identities as soon as possible to not leak unnecessary information.
- Aggregator Level Privacy: The aggregator should preferably learn as little knowledge as possible to do the analytics work. While the quantity of information provided to the aggregator by the events is maintained to a bare minimum, the information subsequently released to the aggregator should fulfill event-level differential privacy. In other words, any numbers presented should be unaffected by the presence or absence of a particular event. Intuitively, this aids in concealing if an event of interest has occurred, such as whether a consumer named Alice has purchased a specific item.

4. Methodology

4.1. Data Preparation Step

As mentioned above, privacy and security are better achieved with context driven processing. In our example, the data is web activity click streams. The number of clusters is predetermined based on the expected size of data. The data are partitioned over K clusters, with n_k number of observations for cluster k, k = 1, ..., K. Let Pk = {1, ..., i, ..., n_k} be the set of indices for k data points; that is, $n_k = |P_k|$. Some more data analysis is performed on available data to determine the correlations/independent variables. This is stored as part of the cluster information in a metadata store. The correlation/independent variable information is used to perform vertical partitions.

4.2. Budgeting and Privacy Allocation

Global privacy loss parameters can be selected based on statistics from PSI's library of differentially private methods. The amount of the global privacy budget that should be allocated for future data analysts will be decided by the data depositor. For example, if the data depositor uses d quanta of privacy for the statistics and chooses to release, each cluster receives an equal budget allocation.

4.3. Event level Privacy

For event level user protection, the data is sanitized at a row level. User-level adjacency is adopted where D and D0 rows are adjacent if D0 can be obtained by adding all the items associated with a single user from Dth row. For each of the rows, the partition is determined based on the cluster score. In accordance with the sensitive and quasi-sensitive qualities, each row is further divided into vertical partitions.

4.4. Aggregator level Privacy

Aggregator privacy is based on the need-to-know principle, which refers to learning just the details necessary to carry out the feature aggregation process. A randomizer method that adds noise is applied to each partition. Only the PRR section of the bloom filters is employed because this is a network-based RAPPOR [17] based encoding, and then the aggregator uses a combination of privacy loss parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

4.5. Budget and Privacy Manager

The choice of differentially private statistics to be computed is made by the data depositor as was explained above. Each partition's vertical partition as well as its management fall under the purview of the privacy manager. It mostly executes two tasks: When a partition's budget has run out, the system 1. monitors its usage and 2. resets it. Within (non-adaptive) batches of queries, the privacy manager employs the fundamental composition Theorem (and its approximations) to compose across several batches.

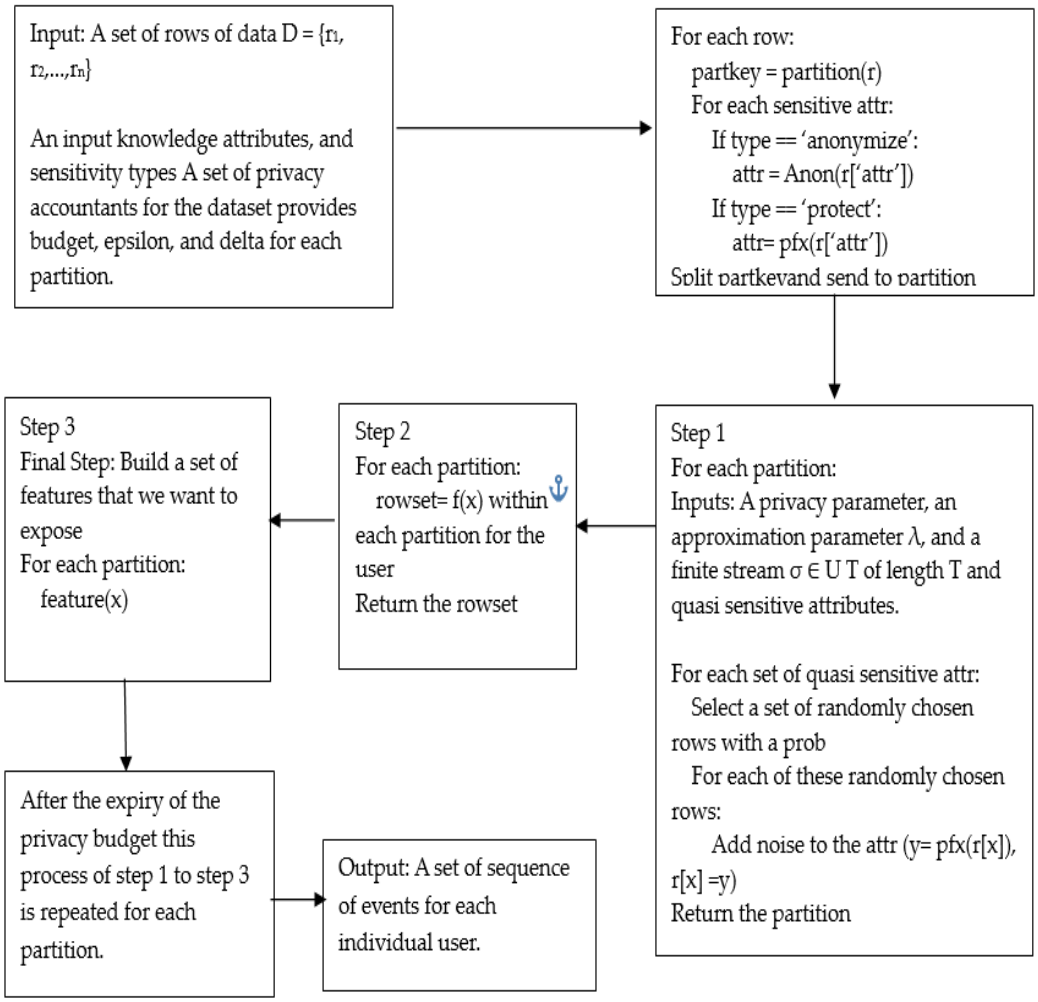

Figure 2. Algorithm for privacy first path analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

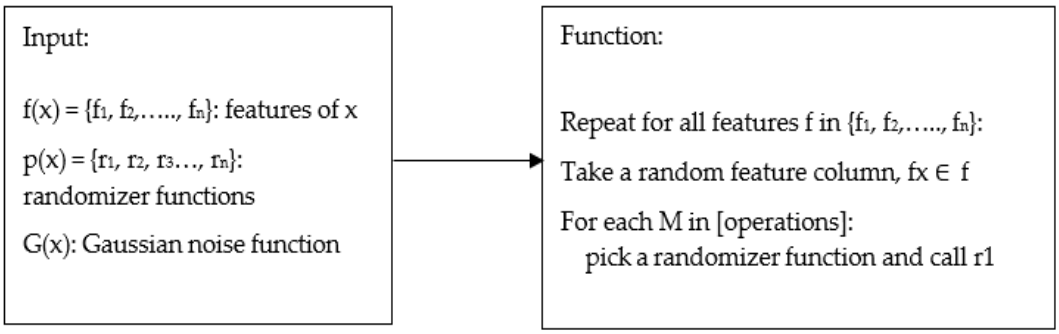

Figure 3. Algorithm for randomization or noise addition: determining the noise to be added

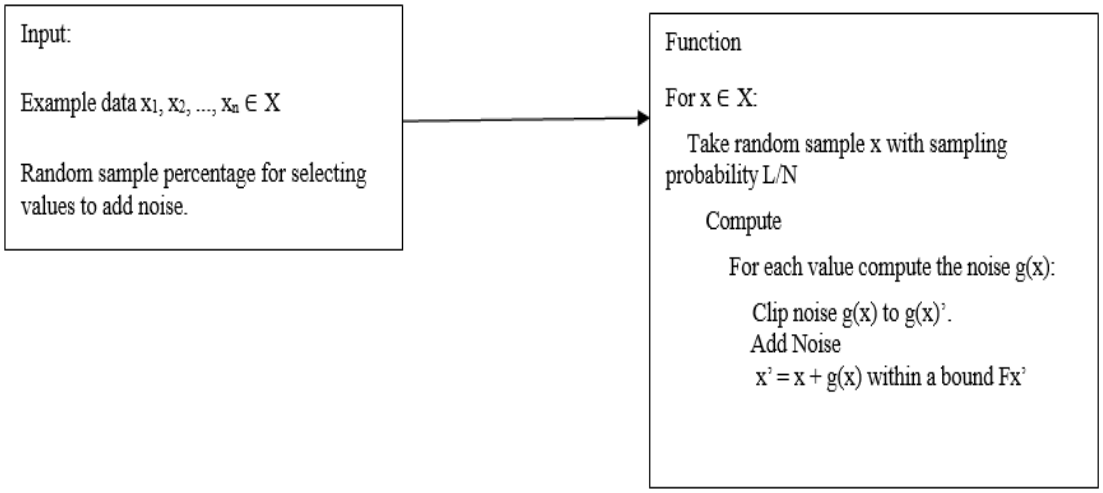

Figure 4. Algorithm for randomization or noise addition: answering a query from a partition

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

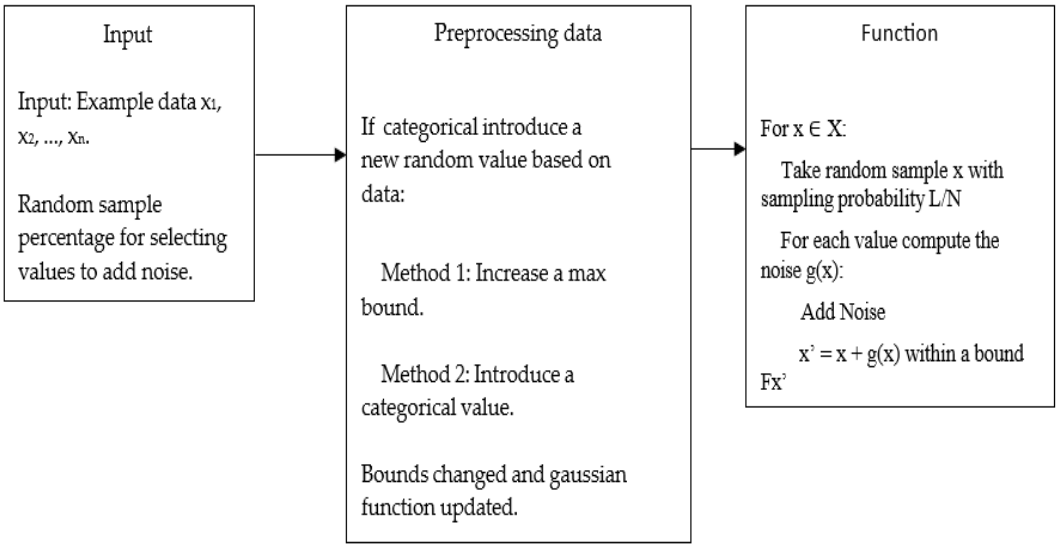

Figure 5. Algorithm for randomization or noise addition: randomizing categorical values

5. Experiments and Results

Clickstream data was used for analysis in the paper. Histograms of the attribute geo_latitude of the first two clusters after noise addition are provided in Figures 6 and 7.

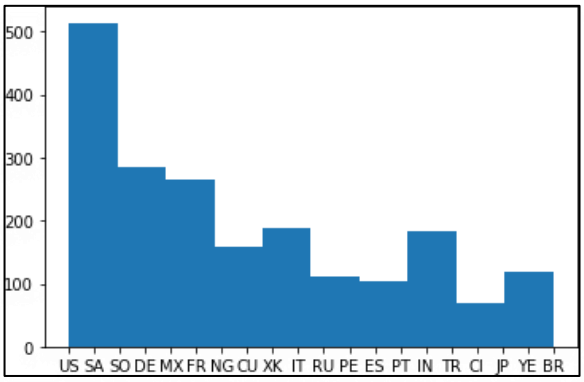

Figure 6. Histogram of the attribute geo_latitude for the first cluster

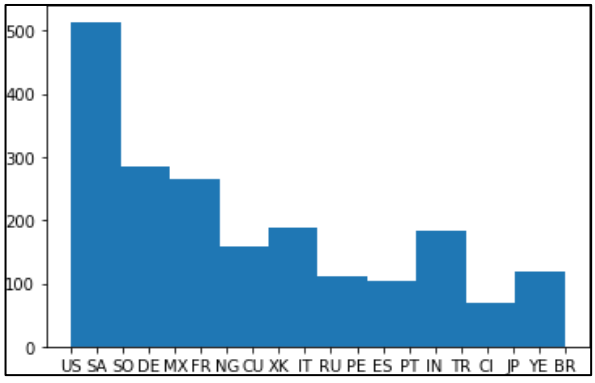

Figure 7. Histogram of the attribute geo_latitude for the second cluster

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

Histograms of the attribute geo_latitude of the first two clusters before noise addition are provided in Figures 8 and 9.

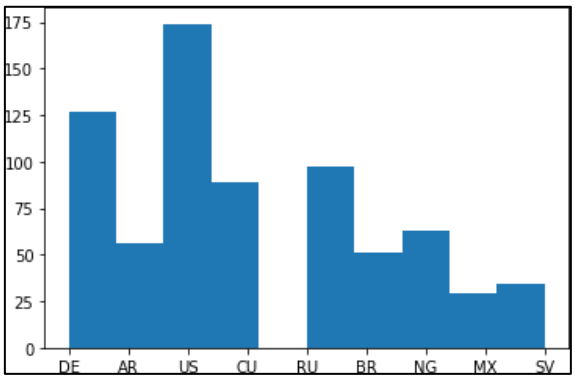

Figure 8. Histogram of the attribute geo_latitude for the first cluster

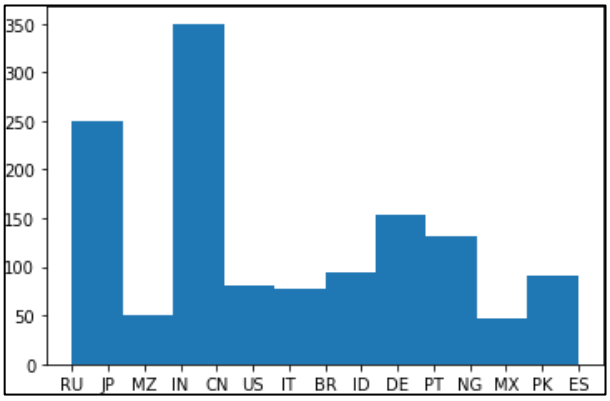

Figure 9. Histogram of the attribute geo_latitude for the second cluster

The Frequency analysis of noisy data and data without noise is shown in Figures 10 and 11.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

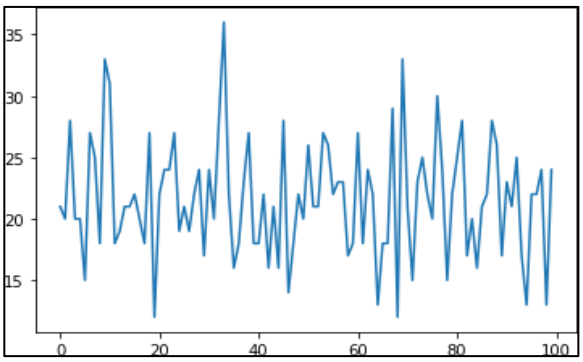

Figure 10. Frequency analysis of noisy data

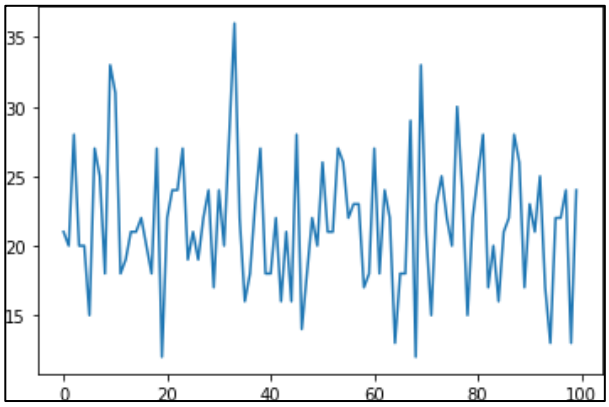

Figure 11. Frequency analysis of data without noise

The Monetary analysis of noisy data and data without noise are as shown in Figures 12 and 13.

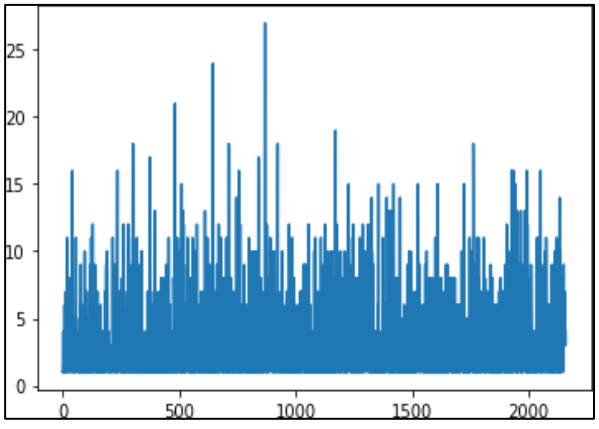

Figure 12. Monetary analysis of noisy data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2023 doi:10.20944/preprints202304.0904.v1

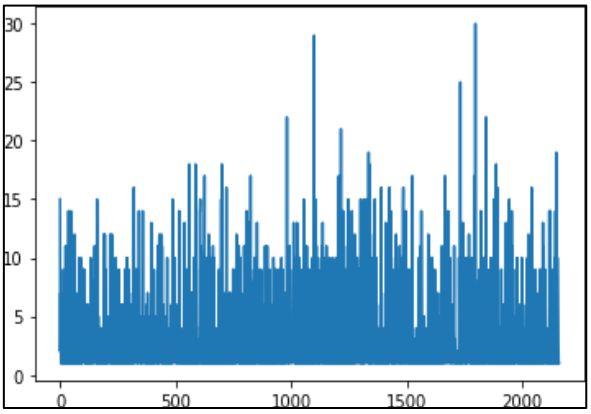

Figure 13. Monetary analysis of data without noise

The Recency analysis of noisy data and data without noise are as shown in Figures 14(a) and 14(b).

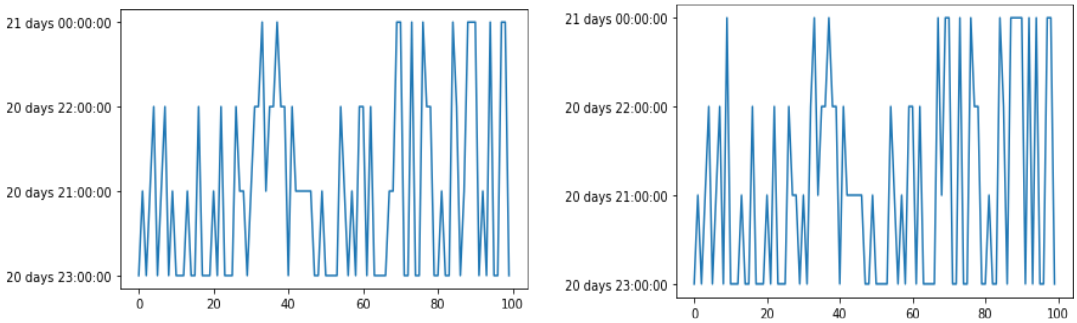

Figure 14(a) and 14(b). Recency analysis of noisy data and Recency analysis of data without noise

6. Conclusion

Data is the holy grail of marketing, consumer engagement, and client retention in today's environment. The data being gathered offers numerous options for businesses to increase revenue, profitability, and customer satisfaction. However, there has been a sharp rise in the amount of data breach events as cloud-based data services have become more widely used. In many nations and states, new laws are constantly being implemented to protect consumer data. Lack of methodology, knowledge, and tools limits these businesses' ability to win over the trust of their clients. Therefore, there needs to be a focused effort made to develop innovative solutions that may aid organizations in making the best use of data. This study focuses on one such strategy that can assist businesses in approaching privacy-aware data stream processing. We utilized a well-known marketing strategy called RFM, which is used to forecast consumer behavior, as an example of this skill. This approach is easily adaptable to other methods as well. By using this application, we demonstrate how businesses can continue to target customers with great precision without having to store, access, or analyze customer data in centralized locations, protecting the private privacy of customers.

Author Contributions: Conceptualization, K.C.G. and S.B.D.; methodology, K.C.G. .; software, K.C.G. and S.B.D.; validation, K.C.G., S.B.D., S.K. and K.N.; formal analysis, K.C.G. and S.B.D.; investigation, S.K. and K.N.; resources, S.K. and K.N.; data curation, K.C.G.; writing—original draft preparation, K.C.G. and S.B.D. ; writing—review and editing, K.C.G., S.B.D., S.K. and K.N.; visualization, K.C.G. and S.B.D.; supervision, K.N.; project administration, K.N.; funding acquisition, S.K. and K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not Applicable

Informed Consent Statement: Not Applicable

Data Availability Statement: Clickstream data | Zenodo

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M. and Michael, J., 2019. Privacy-preserving process mining: Differential privacy for event logs. Business & Information Systems Engineering, 61, pp.595-614.
2. Bagdasaryan, E., Poursaeed, O. and Shmatikov, V., 2019. Differential privacy has disparate impact on model accuracy. Advances in neural information processing systems, 32.
3. Sweeney, L., 2000, August. Foundations of privacy protection from a computer science perspective. In Proceedings of the Joint Statistical Meeting, AAAS.
4. Hann, I.H., Hui, K.L., Lee, T.S. and Png, I.P., 2003. The value of online information privacy: An empirical investigation. Industrial Organization, Econ-WPA.
5. Tucker, C., 2011, May. Social networks, personalized advertising, and perceptions of privacy control. In Proceedings of the tenth workshop on the economics of information security (WEIS).
6. Goldfarb, A. and Tucker, C.E., 2011. Online advertising, behavioral targeting, and privacy. Communications of the ACM, 54(5), pp.25-27.
7. Johnson, G., 2013. The impact of privacy policy on the auction market for online display advertising.
8. Rafieian, O. and Yoganarasimhan, H., 2021. Targeting and privacy in mobile advertising. Marketing Science, 40(2), pp.193-218.
9. Moe, W.W. and Fader, P.S., 2004. Dynamic conversion behavior at e-commerce sites. Management Science, 50(3), pp.326-335.
10. Montgomery, A.L., Li, S., Srinivasan, K. and Liechty, J.C., 2004. Modeling online browsing and path analysis using clickstream data. Marketing science, 23(4), pp.579-595.
11. Park, Y.H. and Fader, P.S., 2004. Modeling browsing behavior at multiple websites. Marketing Science, 23(3), pp.280-303.
12. Hui, S.K., Fader, P.S. and Bradlow, E.T., 2009. Path data in marketing: An integrative framework and prospectus for model building. Marketing Science, 28(2), pp.320-335.
13. Shokri, R. and Shmatikov, V., 2015, October. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1310-1321).
14. Jin, G.Z., 2018. Artificial intelligence and consumer privacy. In The Economics of Artificial Intelligence: An Agenda (pp. 439-462). University of Chicago Press.
15. Paletta, D., Yadron, D. and Valentino-Devries, J., 2015. Cyberwar ignites a new arms race. Wall Street Journal, 11.
16. Dwork, C., 2006. Differential privacy. In Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33 (pp. 1-12). Springer Berlin Heidelberg.
17. Erlingsson, Ú., Pihur, V. and Korolova, A., 2014, November. Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security (pp. 1054-1067).