WHITE PAPER

PRIVATIZED MACHINE LEARNING FOR MARKETING ANALYTICS

FRONTIERS OF MARKETING DATA SCIENCE JOURNAL

neustar®

Joao Natali
Neustar

Robert Stratton
Neustar

Classifications, Key Words:

- privacy preserving machine learning
- differential privacy
- homomorphic encryption
- federated learning

Previously published in Frontiers of Marketing Data Science Journal Issue No. 3 | 2021

Abstract

Two competing trends are shaping the current marketing analytics landscape. On the one hand more and more data is being generated and stored, and on the other privacy regulations and corporate policy threaten the analyst's ability to access and learn from this data. The confluence of these two forces has naturally led to technological innovations that seek to maintain the utility of the granular data for analytical purposes while at the same time offering privacy guarantees to the subjects to whom the data pertains. In this paper we evaluate several privacy preserving technologies in a workflow that reflects a typical real-world marketing analytics deployment. We study the impact of these approaches on computational cost, model fit, attribution accuracy, and the privacy of the simulated individuals, and propose some guidelines for implementing privacy preserving methods in marketing analytics.

1. Introduction

Two competing trends are shaping the current marketing analytics landscape. On the one hand, more and more data is being generated and stored, and on the other hand, privacy regulations and corporate policy threaten the analyst's ability to access and learn from this data. The confluence of these two forces has naturally led to technological innovations that seek to maintain the utility of the granular data for analytical purposes while, at the same time, offering privacy guarantees to the subjects to whom the data pertains.

Data has been called 'the oil' of the digital economy (Wedel & Kannan, 2016). Digitization has led to lower costs of data collection, storage, and transmission (Goldfarb & Tucker, 2019), while rapid growth in media channels, devices, and applications has led to a diverse range of data streams that reflect consumer behaviors, interactions, and responses. This surge in information has provided new opportunities to use data to provide enhanced experiences and satisfaction, while also providing companies with insight into how their advertising efforts are performing at a granular level. These capabilities have had a significant impact on corporate financial performance (Wedel & Kannan, 2016).

But in parallel with this growth in consumer data, questions about data privacy have become increasingly prevalent, with regulations like GDPR and CCPA — that require companies to modify their data handling practices — coming into force. In addition, Wierenga et al. argue that increased sensitivity to privacy concerns has prompted self-policing by many firms, making them reluctant to share data outside of their own firewalls (Wieringa et al., 2021).

While some observers predict that as a consequence of these growing privacy concerns first parties will retain access to their granular data while requiring those on the outside to live with data in aggregated forms, others see potential in an emerging field of privacy-preserving technologies. Although research into privacy-preserving analytics has a long history spanning multiple disciplines, in recent years there has been a significant spike in interest among academics, spawning almost 18,000 papers in 2020 alone. Given that the field is moving so quickly, it can be difficult to extract and implement useable methods from the research, and although the practice of machine learning on granular marketing data falls within the broader scope of existing academic research into private learning, relatively little work has been done to assess the practical implications of privatizing machine learning pipelines for marketing analytics applications.

In this paper, we evaluate several privacy-preserving technologies in a workflow that reflects a typical real-world marketing analytics deployment. To achieve this, we generate a population of individuals, then learn their sensitivities to different marketing stimuli under two privacy-preserving regimes. The first regime assumes that the data itself must be protected using input perturbation. The second takes an algorithmic perturbation approach, applying privacy protection to the machine learning model itself. We study the impact of these approaches on computational cost, model fit, attribution accuracy, and the privacy of the simulated individuals, and propose some guidelines for implementing privacy-preserving methods in marketing analytics.

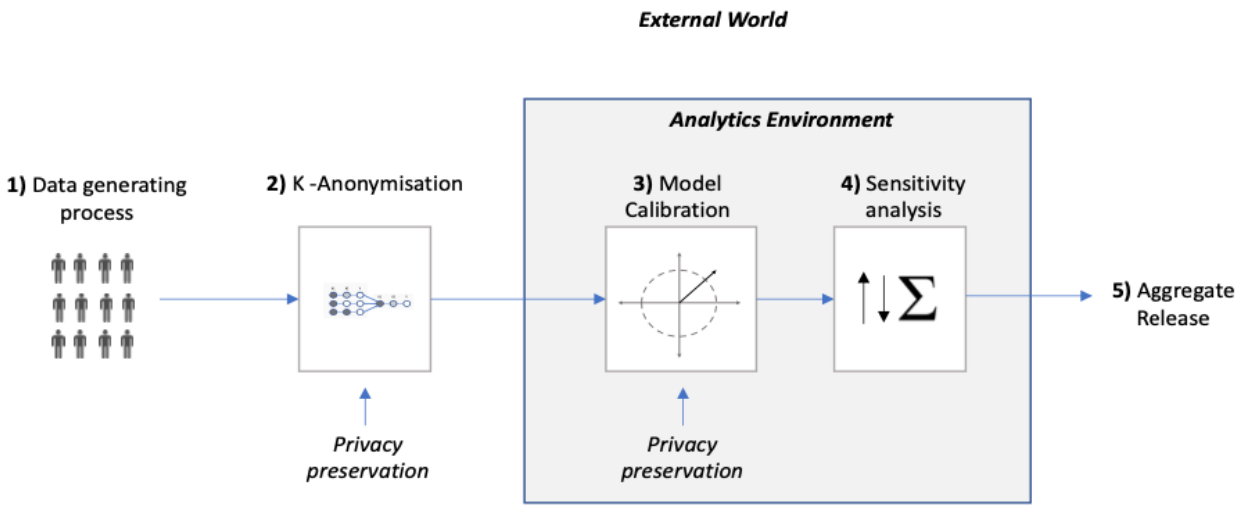

Figure 1: Overview of the simulation environment and the five steps involved in a single end to end run of the process

2. Simulation Environment

We created a simulated environment made up of an external world, in which the agents representing individuals are acting and generating data, and in which the analytics environment is contained. The data is passed from the external world into the analytics environment and progresses through a series of processes, steps 1 through 5 in **Figure 1**, and are detailed further below. At the end of the analytics process, aggregates are created and exported back into the external world. There are two key advantages to using simulated data in this kind of exploration. Firstly, we know the real answers, so we can benchmark the accuracy of different privatized methodologies against non-privatized methods. Secondly, we can examine the impact of counter-factual scenarios, allowing us to look at what would have happened under different conditions.

3. Security Assumptions

The range of threats to which the marketing analytics pipeline is exposed depends to a large extent on the types of interface with the model and data that are available to an attacker. We assume here that the attacker does not have access to the analytics environment itself — that it is secured against intrusion — and that the vulnerabilities are limited to the points in the process that are accessible to the external world — i.e the input data and the aggregated outputs.

Attacks on machine learning pipelines are generally classified into one of three categories, differentiated in terms of the attacker's intentions (Papernot et al., 2016). Attacks on confidentiality aim to recover the model structure or parameters, or the data used to train it. Attacks on integrity seek to induce particular outputs or behaviors of the attacker's choosing. These attacks often involve 'poisoning' the training data to mislead the model about the correct classification for a given input (Jagielski et al., 2018).

Finally, attacks on availability of the pipeline attempt to prevent access to model outputs or other features of the system.

We focused in this study on potential attacks on confidentiality. Because the analytics environment does not expose a scoring API to the external world, we exclude consideration of attacks such as reverse engineering of the training data, model weight and hyperparameter stealing, and membership inference attacks (Shokri et al., 2017).

4. Simulation Process

4.1 Simulation Step 1: Data Generating Process

The synthetic data generating process consists of:

- A heterogeneous population of simulated software agents, each of which represents an individual. Each of the individuals is given a base likelihood of making a transaction, the ability to receive advertising messages, and the ability to make a purchase when they reach a certain utility for the simulated product. The individuals have a variable level of responsiveness to advertising.
- An advertiser with the ability to use two different media channels to deliver advertising impressions to the customer agents over time. Individual agents are exposed to different levels of advertising, reflecting heterogeneity in their underlying media consumption habits.

The simulated data is generated and passed to the k-anonymization system.

4.2 Simulation Step 2: K-anonymization

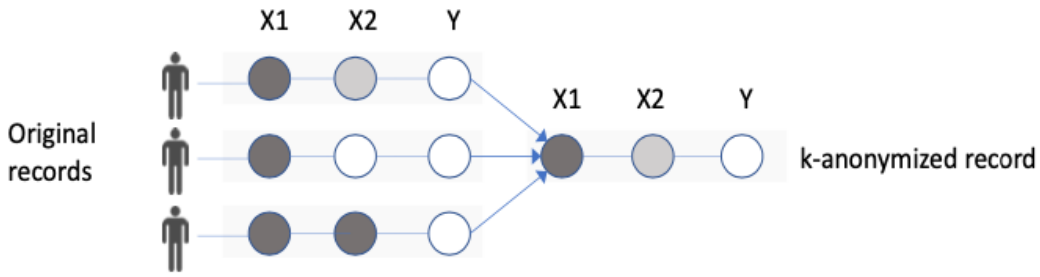

Figure 2: An example of a k-anonymization process where k=3 in which three sets of quasi-identifiers are generalized to a common set

The intuition that motivates **k-anonymization** is that attributes of an individual that uniquely identify them may remain in a record, even after traditional forms of de-identification have been performed, as illustrated in **Figure 2**. For example, in the US, zip code and date of birth create a unique combination of attributes for a large part of the population. An attacker that is unconstrained by any form of process control and has access to these attributes from another source can match them against the record in question and re-identify the individual (Sweeney, 2002).

These attributes are usually referred to in the literature as **quasi-identifiers**. To protect against this kind of re-identification, Sweeney proposed **k**-anonymization as a guarantee that each combination of quasi-identifiers appears "with at least **k** occurrences" in a given dataset. These groups of quasi-identifiers are often referred to as Equivalence Classes (EQs) (Ayala-Rivera et al., 2014).

In this exercise, we use the Mondrian algorithm, a 'greedy' multidimensional approach that recursively partitions the domain space into regions that contain at least k records that share the same EQ.

The data the agents generate, X1, X2, and Y, is collected and assembled, then passed through the Mondrian **k-anonymization** process which may be set at any level of **k**. When k is set to 1, the data is unmodified by the k-anonymization process. Where **k** is > 1, the algorithm generalizes the quasi-identifiers such that EQs satisfy **k** by setting X1, X2, and Y to their mean values.

- Distances between original and k-anonymized quasi-identifiers in the dataset are computed as the L1 norm of the features' differences. Distances between the original (f) and k-anonymized ($\bar{f}$) features are calculated using an averaged L1 norm of the difference between features over the entire dataset (T):

$$d(f,\bar{f}) = \frac{\|f-\bar{f}\|}{|T|} = \frac{1}{|T|}\sum_{i=1}^{|T|}|f_i - \bar{f}_i|$$

- The Discernibility metric represents how indistinguishable a record is from others in the dataset. If a record belongs to an Equivalence Class (EQ) containing $|EQ|$ members, we define the discernibility of the record as $1/|EQ|$. The discernibility of the entire dataset is defined as the sum of the discernibility of each record:

$$D(T) = \sum_{i\in\{EQ\}} \frac{1}{|EQ_i|} = |\{EQ\}|$$

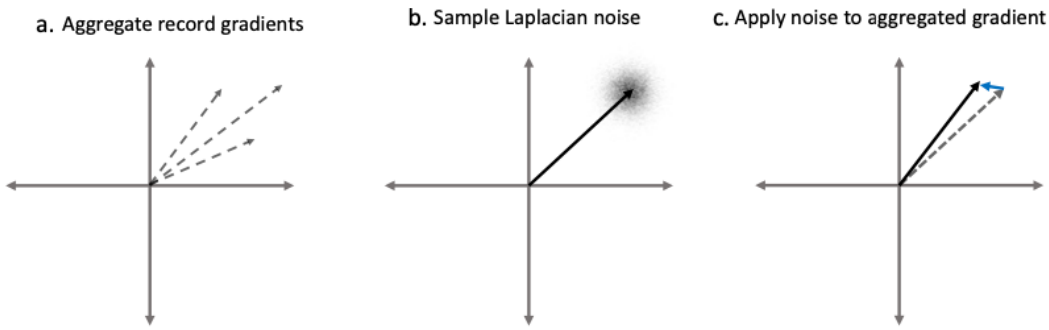

Figure 3: Steps in the computation of the differentially private gradients for the privacy-preserving model estimation. a. Record-level gradients are aggregated into coefficient update direction; b. Laplace noise is sample with a scale proportional to gradient sensitivity and inversely proportional to ε; c. Gradient vector is updated.

4.3. Simulation Step 3: Model Calibration

Some machine learning models and estimation algorithms are susceptible to leaking private information from records used during training by potentially encoding information that would otherwise not be known had an individual not been present in the training dataset. As a simple example, consider a situation in which an individual has a property unseen in other records, but which is nonetheless encoded as a feature in a model. Any association captured through the model estimation process between the individual's property and that their outcome will allow for the identification of an individual with such property in the training dataset.

In this study, we explore the impact of privacy preservation using Logistic Regression models, which are commonly estimated using gradient-based methods. In such estimation algorithms, the encoding of individual-level information into model coefficients occurs through the effect of the gradient of the loss function with respect to a single record exerts in the direction of updates of coefficients at each iteration of the algorithm. A typical strategy to avoid privacy violations through model training is, therefore, to ensure that the loss gradients calculated at each algorithm iteration do not expose information from any single record.

To achieve this goal, we employ a Differentially Private approach to computing loss gradients in our Gradient Descent (GD) based training algorithm, as illustrated in **Figure 3**. At each iteration of the algorithm, we calculate the Jacobian of the record-level log-likelihood loss function with respect to the model parameters:

$$g_{ij} = \frac{\partial L_i}{\partial \beta_j} \quad \forall i \in T, \forall j \in F$$

Where T is the set of all records in the training dataset, F is the set of all features in the model, L_i is the log-likelihood loss for record i, and β_j is the coefficient for feature j.

For each model feature j, we then compute the sensitivity of the loss gradient with respect to j as the smallest value S_j such that for every pair of datasets T and $\check{T}$ differing by a single record,

$$\frac{1}{|T|}\left|\sum_{i \in T} g_{ij} - \sum_{i \in \check{T}} g_{ij}\right| \leq S_j$$

Finally, we compute the loss gradient with respect to each model coefficient as

$$g_j = \frac{1}{|T|}\sum_{i \in T} g_{ij} + \text{Lap}\left(\frac{S_j}{\varepsilon}\right) \qquad \forall j \in F$$

The above mechanism is demonstrated to be ε-indistinguishable by Dwork et al. (2006), and therefore has privacy leakage bounded by ε.The estimation algorithm is then:

- Set learning rate α, set tolerance τ
- Initialize model parameters: $\beta \leftarrow \beta_0$
- while iteration <= max_iterations do:
  - Calculate Jacobian:
    $g_{ij} = \partial L_i / \partial \beta_j$ for every record i and feature j
  - Calculate Sensitivity:
    $S_j = \max(|g_{ij}|)$ for every feature j
  - Calculate gradients: $g_j = \sum_i g_{ij} + \mathrm{Lap}(S_j/\varepsilon)$ for every feature j
  - Update model parameters: $\beta_j \leftarrow \beta_j - \alpha \cdot g_j$ for every feature j
  - if $\sqrt{\sum_j g_j^2} \le \tau$
    Stop
    Otherwise
    iteration = iteration + 1

4.4 Simulation Step 4: Sensitivity Analysis and Aggregation

Once we have a calibrated model, we conduct a sensitivity analysis using the model and the dataset together to assess the contribution of each of the media channels, X1 and X2, and the base level of sales.

We then sum the contributions of each of the channels.

- Calculate the total sales available in the entire dataset

$$total\_sales = \sum_i^I \sum_t^T y_{it}$$

- Calculate the total sales with X1 excluded

$$attributon\_to\_all\_exept\_X1 = \sum_i^I \sum_t^T \left(\frac{1}{(1+\exp(-(\beta 1 + \sum_{t=0}^{t} X1_{it} * 0 + \sum_{t=0}^{t} X2_{it} * \beta 3)))} \right)$$

- Calculate the total sales with X2 excluded

$$attributon\_to\_all\_exept\_X2 = \sum_i^I \sum_t^T \left(\frac{1}{(1+\exp(-(\beta 1 + \sum_{t=0}^{t} X1_{it} * \beta 2 + \sum_{t=0}^{t} X2_{it} * 0)))} \right)$$

- Calculate sales attributable to base

$$attributon\_to\_base = \sum_i^I \sum_t^T \left(\frac{1}{(1+\exp(-(\beta 1 + \sum_{t=0}^{t} X1_{it} * 0 + \sum_{t=0}^{t} X2_{it} * 0)))} \right)$$

- Calculate sales attributable to X1:
  - total_sales – attribution_to_all_except_X1
- Calculate sales attributable to X2:
  - total_sales – attribution_to_all_except_X2

4.5 Simulation Step 5: Aggregate Release

The aggregate results produced in Step 4 are the output from the analytics environment into the external world.

5. Simulation Results

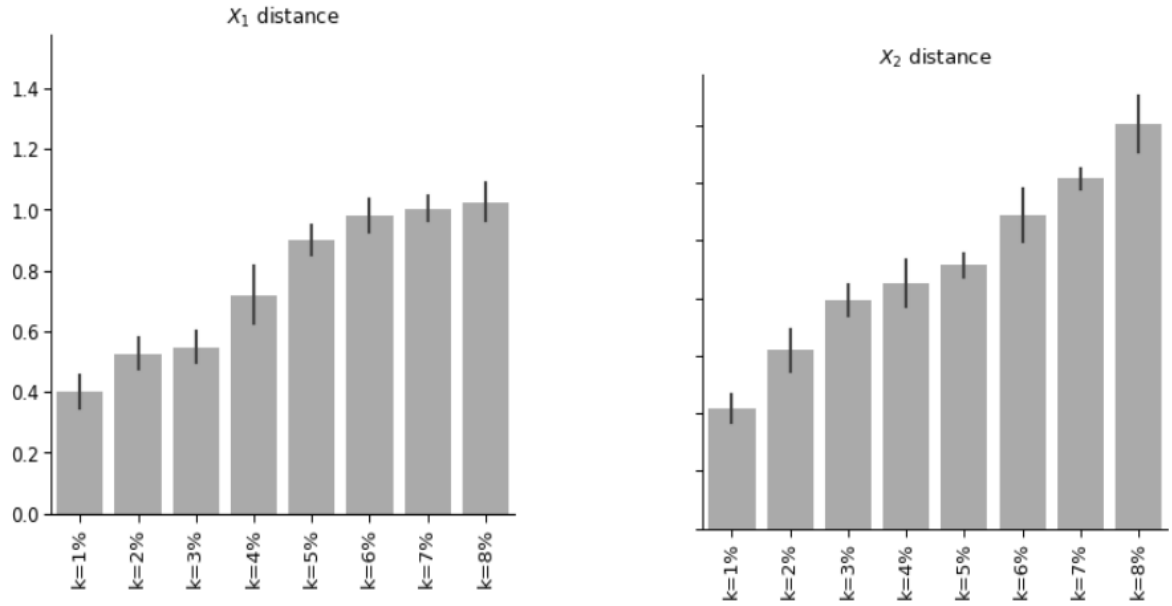

Figure 4: Distances between original and *k*-anonymized quasi-identifiers in the dataset, computed as the L1 norm of the feature differences. Thin bars represent the confidence interval of the measured value, defined as one standard deviation above and below the mean.

As would be expected, **Figure 4** shows that as the level of ***k*** increases, the average distance between each X1 and X2 value and its ***k***-anonymized counterpart increases. Even at high levels of ***k*** though, frequently occurring combinations of values may be sufficiently common that no anonymization is required since their pre-existing EQ may already meet the ***k*** requirement. In parallel, as ***k*** increases the discernibility of any record in the dataset decreases (see **Figure 5**). As **Figure 6** shows, the computational cost of applying ***k***-anonymization for low values of ***k*** can be substantial. As ***k*** increases, the number of partitions the Mondrian algorithm needs to apply to find EQs reduces.

For the purposes of this study, we assumed that combinations of X1, X2, and Y may act as quasi-identifiers that can potentially be used to uniquely identify the simulated individuals in the dataset. In the real world, it may be more likely that only a subset of these quasi-identifiers may be accessible to an attacker.

For the particular characteristics of the data generating process that we were using, the impact of ***k***-anonymization on AUROC was negligible until ***k*** approaches 8% of the total number of records in the dataset (see **Figure 7**). There may be no direct comparison of these percentages to other datasets though, since the distributions of the variables and their

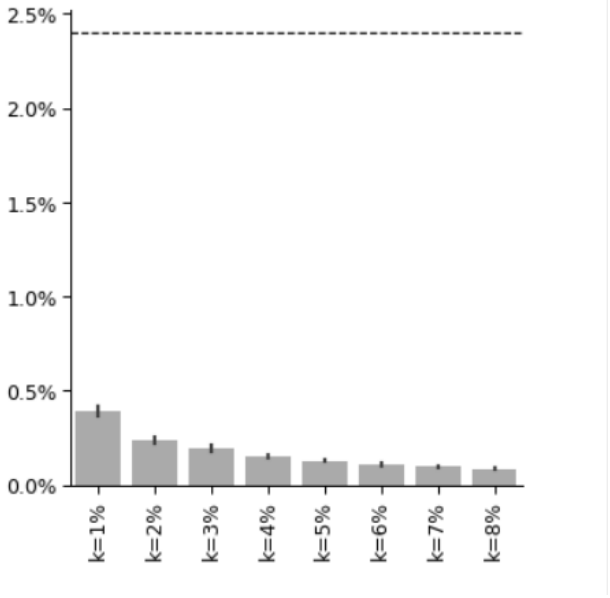

Figure 5: Ability of an attacker to single out an individual from the set of observations for different values of *k* as a fraction of the size of the dataset. Thin bars span one standard deviation above and below the mean.

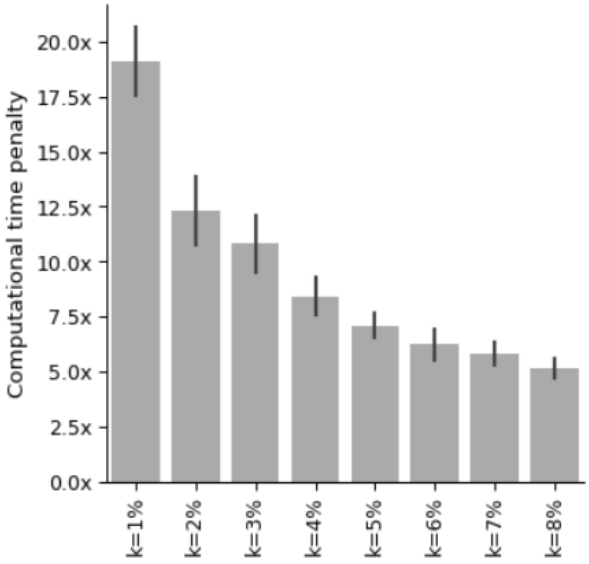

Figure 6: Variation of computational time in relation to a non-anonymized model estimation observed by applying the Mondrian *k*-Anonymity, for different values of *k* as a fraction of the size of the dataset. ε was set to a large value to avoid noise addition to the estimation.

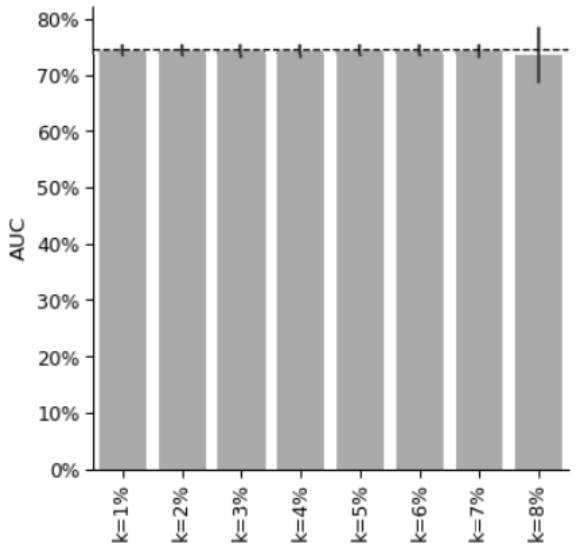

Figure 7: Model predictive power measured as the area under the ROC curve (AUROC), for different values of k as a fraction of the size of the dataset. The dotted line represents the AUROC obtained for the model estimated on the non k-anonymized dataset. ε was set to a large value.

inter-relationships will play a major role. Also, as **Figure 8** shows, at reasonable practical levels of ***k*** we observed minimal bias in the attribution results. For example, where ***k*** =1% of total records, e.g. a relatively large ***k*** value of 10,000 in a million-record dataset, the bias in attribution is only 1%. As ***k*** increases, though, the level of bias trends upwards, and there would be limits on the practical usefulness of machine learning results from ***k***-anonymized data at some level of ***k***. In the data generating process and learning pipeline we used, attribution is generally

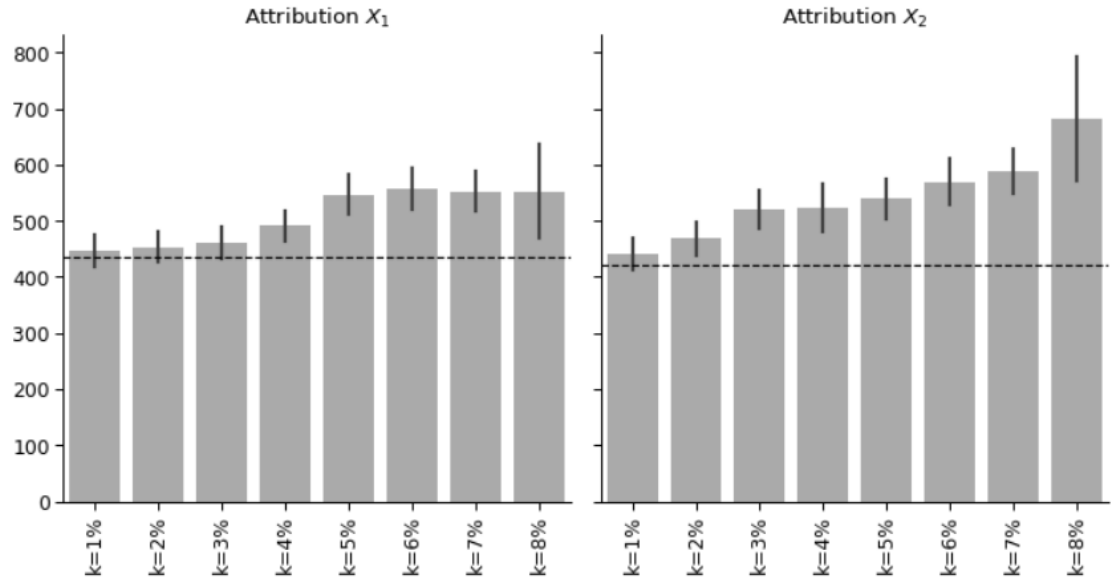

Figure 8: Outcome attributed to each model driver for different values of *k* as a fraction of the size of the dataset. The dotted line represents the attribution obtained without applying *k*-anonymity. Thin bars span one standard deviation above and below the mean. ε was set to a large value to avoid noise addition to the estimation.

biased upwards as ***k*** increases, but it is not clear that the bias would always be in this direction for any dataset.

In experiments that pushed ***k*** beyond the values reported here, we discovered a number of 'breaking points' that make ***k***-anonymization above a certain level of ***k*** impractical. For example, if ***k*** exceeds the number of outcomes of a particular class in a dataset, the outcome cannot be preserved in a 1/0 form but itself becomes the average value of multiple classes.

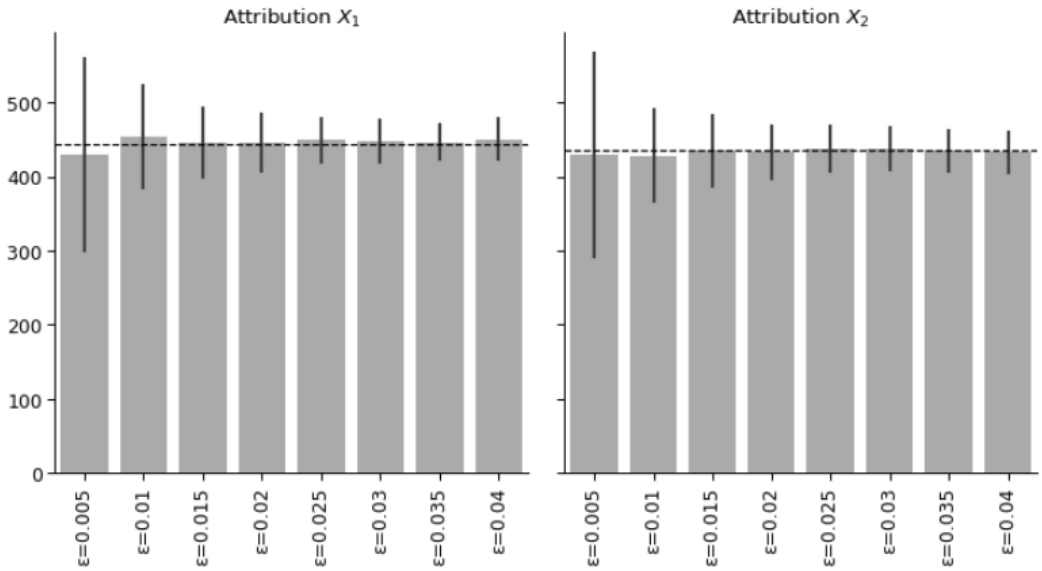

Figure 9: Outcome attributed to each model driver for different values of ε. Thin bars span one standard deviation above and below the mean. The dotted line represents the attribution obtained with a model estimated without privatized learning. k value was set to 1.

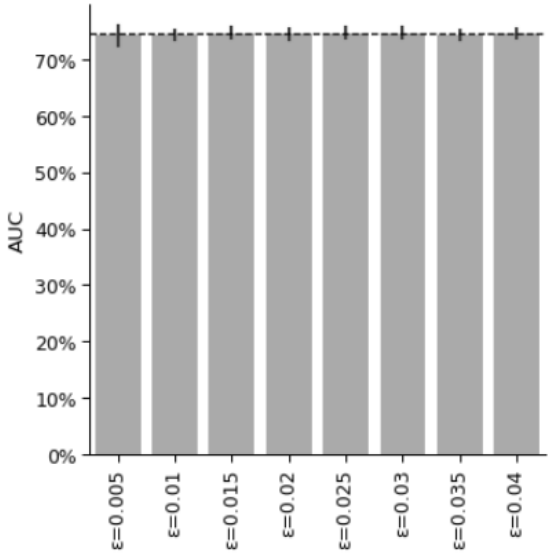

Figure 10: Distances between original and k-anonymized quasi-identifiers in the dataset, computed as the L1 norm of the features differences. k value was set to 1.

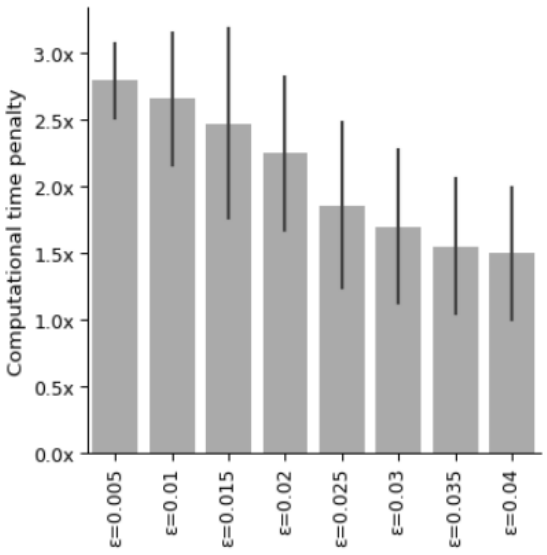

Figure 11: Variation of computational time in relation to a non-anonymized model estimation observed for different values of ε. Thin bars span one standard deviation above and below the mean. k value was set to 1.

In general, differential privacy can be achieved by adding a reasonable amount of noise into the output results of a computation over a dataset. The amount of noise will affect the trade-off between privacy and utility of the results. Specifically, too much noise will make the dataset useless and too little noise (Gong et al., 2020) is insufficient to provide privacy guarantees. In our model estimation approach, this trade-off is modulated by the value of selected.

The chosen value of ε specifies within each Gradient Descent iteration, a bound on the ratio of probabilities of calculating a specific gradient component value over two datasets that differ only by a single record. Therefore, it limits the impact that any record can have in the gradient computation, preventing the model from encoding information particular to any single observation into the model parameters.

In practice, smaller values of ε yield stronger privacy protection, at the cost of higher added variation in calculated gradients and in attribution results (**Figure 9**). This variation may impair the effectiveness of the parameter search algorithm and result in higher computational costs for training, as observed in **Figure 11**, while not impacting model accuracy (**Figure 10**).

6. Towards an integrated view of privacy

In the previous sections, we explored the practical impact of applying two types of privacy preservation in a machine learning pipeline. To understand the role that these mechanisms play individually and their potential to work together, it is worth stepping back and reviewing the different notions of privacy that each supports. The current literature on privacy-preserving technologies does not coalesce around a universal definition of privacy. Some of the more general ideas that motivate research in the area include the suggestions that it is the "right to be let alone" (Warren & Brandeis, 1890), or "protection from being brought to the attention of others." (Gavison, 1980). As technology has enabled the collection of increasingly detailed data about individuals, Dwork argues that "the need increases for a robust, meaningful, and mathematically rigorous definition of privacy" (Dwork & Roth, 2013). ***K***-anonymity and differentially private gradient descent are two examples of mathematically rigorous definitions of privacy, but they are not competing approaches to achieve the same outcome, and each was developed for a different purpose with a different idea of privacy in mind.

K-anonymity was conceived as a method for protecting the identities of individuals in published data, for example, people whose details are contained in the release of medical data from a hospital. As such, it applies privacy protection to the data itself and has an impact on any analysis or query that is done on it. It guarantees that each entity contained in the data cannot be distinguished from at least k individuals whose attributes also appear in it.

Differential privacy was designed to protect the privacy of individuals by applying noise to the results of aggregate queries in which they may be present, but does not rely on transformations on the underlying data itself. Instead, it guarantees that an algorithm's ***output*** does not differ significantly statistically for two versions of the data differing by only one record.

K-anonymization and privacy-preserving learning, therefore, offer two different types of privacy protection whose applicability is determined by the constraints we impose in the measurement system. The former allows for strong guarantees around the limitations of an attacker's ability to discern a user from others in the dataset, whereas the latter offers well-defined limitations on the impact a single user's data have on the learning of the measurement model.

As our simulations have shown, these two approaches also have starkly distinct effects on both the increment in computational costs

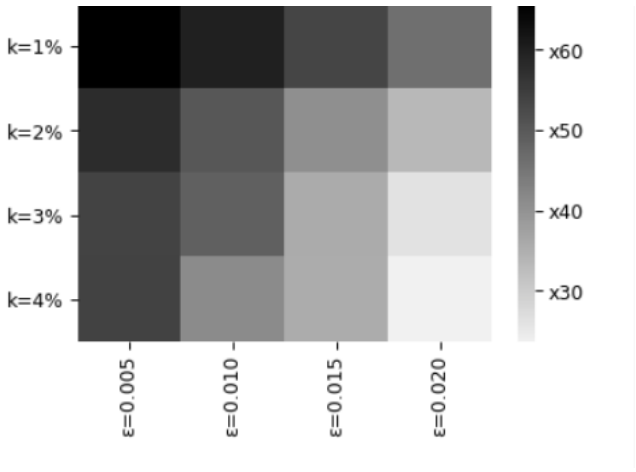

Figure 12: Computational penalty due to the sequential application of k-anonymization and privacy-preserving learning, as a factor over the baseline non-private model estimation.

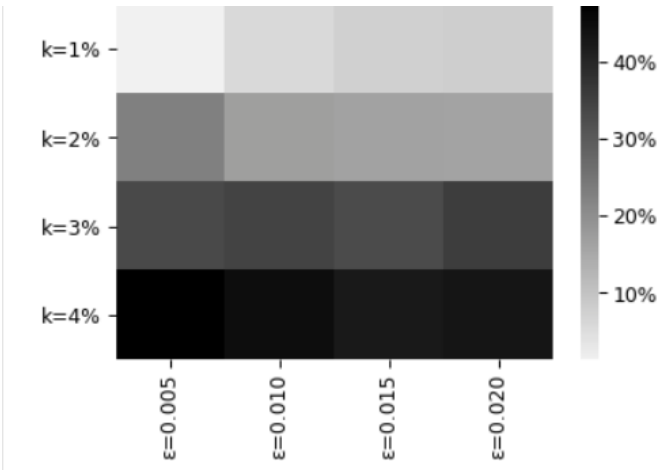

Figure 13: Mean of attribution error resulting from the application of k-anonymization and privacy-preserving learning. Higher values imply a larger bias on attribution error.

and accuracy of the attribution measurement. In **Figure 12**, we can see that stronger ***k***-anonymity constraints (larger values of ***k***) result in lesser computational penalty added to the measurement process, since a smaller number of partitions of the dataset are required to meet the ***k*** constraint. Conversely, stronger privacy preservation at the level of model training results in higher computational costs, due to the increased difficulty of finding a solution to the estimation problem under higher amounts of added noise.

When it comes to measurement accuracy, **Figure 13** shows that ***k***-anonymity adds a stronger bias on the mean attribution calculated under its privacy protection. This bias, however, is only observed at ***k*** values representing higher fractions of the entire dataset and, therefore, is of little practical concern when working with larger datasets. But as we saw in **Figure 9**, privacy-preserving model training does not add any significant bias to attribution measurement, although it does add variance to the results due to the noise injected into the model parameter search process.

Conclusion

We have studied the application and effects of two popular privacy preservation techniques to the problem of measuring marketing impact under the controlled environment of simulated data. Our analysis shows that there are a number of considerations a marketing measurement practitioner must make in order to effectively apply privacy preservation on an attribution system. ***K***-anonymity and private learning offer distinct sets of trade-offs related to their impact on measurement accuracy, results variance, and computational costs. We have found that ***k***-anonymizing process inputs potentially yield measurement biases at high thresholds of ***k***, whereas lower ***k*** values result in super-linear increases in computational costs. Conversely, privacy-preserving model estimation does not result in a significant measurement bias, even at very low values of ε, but will increase the variance associated with measurement, as well as training costs, at higher strengths of privacy preservation.

It is worth noting that privacy preservation techniques do not come without computational cost. As shown in **Figures 6** and **11**, which are expressed as a penalty with respect to an end-to-end model estimation process that does not include a privacy-preservation step, there is a time penalty associated with each of the methods that we evaluated.

Ultimately, the choice of which privacy protection method to employ — or combination of methods — and which protection strength lies on the balance between the measurement accuracy considered acceptable, the computational costs involved in the process, and the levels of privacy protection deemed necessary to be imposed on the system. Our results indicate that using a ***k*** below 1% of the dataset size yields very small biases and may still be computationally feasible for most use-cases. A value of greater than 0.02, still considered small for many applications, results in a low impact in both accuracy and computational costs.

References

1. Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing, 80*(6), 97–121. Source

2. Goldfarb, Avi, & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature, 57*(1), 3–43. Source

3. Wieringa, J., Kannan, P., Ma, X., Reutterer, T., Risselada, H., & Skiera, B. (2021). Data analytics in a privacy-concerned world. *Journal of Business Research, 122*, 915–925. Source

4. Papernot, N., Mcdaniel, P., Sinha, A., & Wellman, M.P. (2016). Towards the Science of Security and Privacy in Machine Learning. ArXiv, abs/1611.03814.

5. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li. B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *2018 IEEE Symposium on Security and Privacy (SP)*, 19–35. IEEE (2018). Source

6. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP), 3–18*. IEEE (2017).

7. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. Source

8. Ayala-Rivera, V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners. *Transactions on Data Privacy 7*(3), 337-370.

9. Gong, M., Xie, Y., Pan, K., Feng, K., & Qin, A. (2020). A Survey on Differentially Private Machine Learning [Review Article]. *IEEE Computational Intelligence Magazine, 15*(2), 49–64. Source

10. Warren, S., & Brandeis, L. (1890). The Right to Privacy. *Harvard Law Review, 4*(5), 193–220. Source

11. Gavison, R. (1980). Privacy and the Limits of Law. *The Yale Law Journal, 89*(3), 421–471. Source

12. Dwork, C., & Roth, A. (2013). *The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science, 9*(3–4), 211–407. Source

13. Bleier, A., Goldfarb, A., & Tucker, C. (2020). Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing, 37*(3), 466–480. Source

14. LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Mondrian Multidimensional K-Anonymity. *22nd International Conference on Data Engineering (ICDE'06)*. Source

15. Li, N., Qardaji, W.H., & Su, D. (2011). Provably private data anonymization: or, k-anonymity meets differential privacy. CoRR, abs/1101.2604, 49, 55 (2011) 15.

16. Song, L., & Mittal, P. (2020). Systematic Evaluation of Privacy Risks of Machine Learning Models. ArXiv, abs/2003.10595.

17. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–19. Source

Authors

Joao Natali is Sr. Director, Data Science, leading analytics R&D for Neustar Marketing Solutions. Joao has over 15 years of experience in data science and marketing analytics, and stays excited by new challenges in the field of technology. He has a Ph.D in engineering, focused on optimization and machine learning applied to full genomic analysis.

joao.natali@team.neustar

Robert Stratton is SVP, Data Science, leading R&D efforts across Neustar Marketing Solutions. Robert has over 15 years of experience leading and conducting analytics projects across a wide range of industries and applications, from digital attribution and transactional analysis to process mining, marketing mix and brand equity analysis. He has a PhD from King's College in London and is an expert on computational modeling.

robert.stratton@team.neustar

ABOUT NEUSTAR

Neustar is an information services and technology company and a leader in identity resolution providing the data and technology that enables trusted connections between companies and people at the moments that matter most. Neustar offers industry-leading solutions in Marketing, Risk, Communications, and Security that responsibly connect data on people, devices and locations, continuously corroborated through billions of transactions. Neustar serves more than 8,000 clients worldwide, including 60 of the Fortune 100. Learn how your company can benefit from the power of trusted connections at www.home.neustar.

WP-MS-335772-08.20.2021