The Projection-Based Data Transformation Approach for Privacy Preservation in Data Mining

Diana Judith Irudaya Raj
Department of Computer Science, Stella Maris College, Chennai, India
dianajudith@stellamariscollege.edu.in

Vijay Sai Radhakrishnan
School of Computing, SASTRA Deemed to be University, India
vijaysai@it.sastra.edu

Manyam Rajasekhar Reddy
School of Computing, Amirta Vishwa Vidyapeetham, Amaravathi Campus, India
m_rajasekharreddy@av.amrita.edu

Natarajan Senthil Selvan
School of Computing, SASTRA Deemed to be University, India
senthilselvan@cse.sastra.edu

Balasubramanian Elangovan
Department of CSE, Koneru Lakshmaiah Education Foundation, India
elan77@gmail.com

Manikandan Ganesan
School of Computing, SASTRA Deemed to be University, India
manikandan@it.sastra.edu (corresponding author)

Received: 29 May 2024 | Revised: 17 June 2024 | Accepted: 24 June 2024

ABSTRACT

Data mining is vital in analyzing large volumes of data to extract functional patterns and knowledge hidden within the data. Data mining has practical applications in various scientific areas, such as social networks, healthcare, and finance. It is important to note that data mining also raises ethical concerns and privacy considerations. Organizations must handle data responsibly, ensuring compliance with legal and ethical guidelines. Privacy-Preserving Data Mining (PPDM) refers to conducting data mining tasks while protecting the privacy of sensitive data. PPDM techniques aim to strike a balance between privacy protection and data utility. By employing PPDM techniques, organizations can perform safe and private data analysis, protecting sensitive information while deriving valuable insights from the data. The current paper uses geometric transformation-based projection techniques such as perspective projection, isometric projection, cabinet projection, and cavalier projection to protect data privacy and improve data utility. The suggested technique's performance was assessed with the K-means clustering technique. The UCI repository's Bank Marketing dataset was used to verify the error rate of the proposed projection techniques.

Keywords-data privacy; data mining; data transformation; projection; k-means algorithm

I. INTRODUCTION

Data is a key component in the field of information technology. Organizations use cloud and grid storage, among other different types of storage, to store data. Numerous organizations utilize many data mining techniques to facilitate strategic decision-making. Numerous disciplines rely heavily on data mining, including image analysis, weather forecasting, and medical diagnosis. A significant drawback of data mining is that some private information is also revealed during the process of mining. Various approaches to deal with this privacy issue have been suggested [1]. The purpose of privacy-preserving techniques is to help data miners without disclosing any extra information that would violate users' privacy. Sensitive attributes are typically suppressed before providing the data to the data miner. Yet, it is still feasible to deduce sensitive information using additional characteristics referred to as quasi-identifier attributes [2]. A sanitization process is used to alter the original database's data to stop privacy leaks. The behavior of the cleaned data ought to be identical to that of the original data. Creating modified data that retains the original data dimensions is the primary goal of privacy-preserving algorithms. There are methods, such as secure multi-party computation, randomization, and anonymization, for ensuring privacy [3, 4]. By introducing random noise before sharing, randomization modifies the data. Generalization and suppression are used in k-anonymization to transform the data. In generalization, a value is substituted with a less precise but semantically sound value. Dimensionality is reduced through suppression, which removes some rows or columns [5].

Protecting sensitive information from misuse by hiding it before it is published is one of the main objectives of privacy protection [6, 7]. For example, a hospital may make patient records available so that researchers can study the characteristics of various diseases. To preserve people's privacy, sensitive information about specific individuals found in raw data is not disclosed. Other attributes present in external data can be used to ascertain an individual's identity. Table I illustrates hospital data published after the sensitive attributes have been suppressed. Table II shows a sample of data from a public voter registration list. If an adversary has access to this data, he can quickly identify all of the patient's information by comparing the two tables using pseudo-sensitive attributes such as zip code, age, and gender. Four types of projection-based data transformation approaches are used in this work's suggested method to guarantee data mining privacy. Validating the proposed model's results is done with the use of the K-means clustering algorithm. Counting the number of data points in the clusters allows one to gauge how effective the suggested method is. Clusters in the original dataset should match those in the modified dataset after the data has undergone transformation. After data transformation, there are occasionally possible issues that could cause a point from one cluster to move to the other. A misclassification error occurs when data points in a distorted dataset are not correctly classified.

The following are the study's primary contributions:

- The proposed strategy ensures data privacy in data mining through projection-based data transformation.
- A comparative analysis was conducted using various projection techniques, including perspective, isometric, cabinet, and cavalier projections.
- The proposed model's effectiveness is compared with existing techniques using the misclassification error rate.
- With the help of the suggested model, data owners can share their information with the research community without worrying about privacy violations or data leaks.

TABLE I. RAW DATA

ID	Age	Gender	Zip Code	Disease
501	22	F	713001	Fever
602	33	M	813002	Tuberculosis
703	55	M	513003	Fever
804	66	F	413004	Jaundice
905	77	F	313005	Blood cancer

TABLE II. VOTER REGISTRATION LIST

ID	Age	Gender	Zip Code	Name
20	22	F	713001	Jack
21	33	M	813002	Jill
32	55	M	513003	Alice
43	66	F	413004	Bob
64	77	F	313005	Tom

II. RELATED WORK

Depending on the sensitivity of the sensitive attribute values, the proposed algorithms (MBF-MSLF, MSDCF-MSLF, and MMDCF-MSLF) offer varying degrees of protection, resulting in superior privacy protection compared to classical L-diversity. Data loss and runtime are two areas where the algorithms' performance is observed to be better based on the experimental results. In [8], the author address the issue of privacy-preserving data publishing in the context of big data, which is becoming more and more pertinent as sensitive data is gathered and examined in more significant quantities. Sensitivity levels are a novel and potentially valuable addition to the anonymization process. The suggested method is scalable and possibly efficient. The possible influence of anonymization on data utility is not discussed.

FPDP (Flexible Privacy-Preserving Data Publishing) offers a fix for data sharing and privacy in balance for astute farming. The ElGamal cryptosystem protects data privacy. Users are provided with options for changing data aggregation techniques and privacy levels. The study [9] primarily examines numerical data, but more research is necessary because real-world smart agriculture data may be more varied. As an innovative and possibly effective replacement for current techniques, vectorization can be used for privacy-preserving data publishing. This work [10] focuses on using temporal and spatial data. New research directions become possible when data co-occurrence analysis is enabled while maintaining privacy. Large-scale datasets may require additional research because the suggested method could be computationally costly. Value swapping [11] is a promising method for publishing data while maintaining privacy, especially for numerical data and applications that call for quick and easy methods. To improve the method's usefulness and applicability, more study and development must be done in the areas of ensuring data utility, reducing the risk of information leakage, and extending privacy guarantees. The KC-Slice model [12] offers flexibility for evolving requirements by enabling dynamic updates to the data without jeopardizing privacy guarantees. Regardless of the differences in sensitivity between each sensitive attribute, KC-Slice applies the same threshold to them all. Inadequate privacy protection for specific attributes or needless data suppression could result from this. More research is necessary to determine whether the findings apply to other than numerical data types, such as textual or categorical data. Authors in [13] examine and discuss various PPDM and PPML techniques, such as federated learning, homomorphic encryption, and secure multi-party computation. There is a lack of a comprehensive assessment of the efficacy of the various techniques and applications presented in this work in actual healthcare settings. This research focuses primarily on technological solutions to privacy issues, with less attention paid to the moral and legal issues pertaining to healthcare data privacy. Using k-anonymization and a hybridized optimization algorithm, authors in [14] presented a privacy-preserving method for publishing big data in the cloud. A well-established framework for privacy protection can be obtained by utilizing k-anonymity, which guarantees that each person's data record is identical to at least k-1 other records. When genetic and differential evolution algorithms are combined, as opposed to being used separately, the search for the best anonymization solutions may be conducted more quickly and effectively. Only artificial datasets were used for evaluation, and there are no comparisons between the experimental results and other cutting-edge privacy-preserving schemes.

Authors in [15] proposed an improved form of l-diversity that guarantees that each equivalence class contains at least l distinct sensitive values. This version of l-diversity addresses the shortcomings of the original l-diversity model. The suggested method makes efficient use of the MapReduce programming paradigm, which makes it appropriate for handling large amounts of data. In comparison to other scalable l-diversity algorithms, the experimental results show that the suggested strategy can achieve superior information loss reduction. A unique method is suggested [16] to lessen the risk of location injection attacks, which are a significant worry for Location-Based Services (LBS). To guarantee that the user's location cannot be distinguished from at least k-1 other users, the suggested method makes use of k-anonymity, a well-established privacy model. The approach in [17] adapts to the changing user location and surroundings, providing continuous privacy protection during continuous LBS queries. Although k-anonymity protects against identity disclosure, it is susceptible to other types of attacks, such as background knowledge and attribute disclosure. The Efficient Data Anonymization Model Selector for Privacy-Preserving Data Publishing (EDAMS) addresses the challenge of balancing data utility and privacy in a world of increasing data collection. By choosing an anonymization model automatically, EDAMS has the potential to save time and money. The speed at which EDAMS anonymizes data is its primary benefit. Only the k-anonymity, l-diversity, and t-closeness models are supported by the current implementation. While efficiency is discussed, it is not explained in detail in [18] how EDAMS accomplishes this. Adaptive thresholding [19] aims to save storage space by only recording significant events. Adaptive thresholding motion detection has a potential for storage optimization, but practical application requires careful consideration of factors such as computational demands, storage efficiency, adaptability, integration complexity, and comparative effectiveness with other technologies.

The primary concern with any encryption technique is its resistance to attacks. Validating the proposed encryption technique in [20] for color images against other technologies is challenging without specific technical details and evidence to support claims of efficiency and high security. Preserving data mining and machine learning applications in healthcare presents numerous challenges, including privacy, data quality, regulatory compliance, and ethical considerations. To address these challenges, healthcare data mining initiatives must follow interdisciplinary frameworks that prioritize patient care, privacy, and ethics. Data mining techniques that protect sensitive information and provide valuable insights have great potential for secure data publishing. Authors in [21] focus on addressing challenges such as privacy-utility trade-offs, scalability, data quality, interpretability, regulatory compliance, and sustainability to advance the field. Although blockchain technology is secure and decentralized, it has limitations in scalability, including transaction processing speed and data storage [22]. Maintaining privacy and confidentiality of sensor data on a public blockchain network presents significant challenges.

II. THE PROPOSED SYSTEM

The proposed approach makes use of the Bank Marketing data set, a standard benchmark dataset that can be found in the UCI Machine repository [23]. The workflow of the proposed system is shown in Figure 1. Our experiment has demonstrated that applying the k-means algorithm to the transformed data increases privacy.

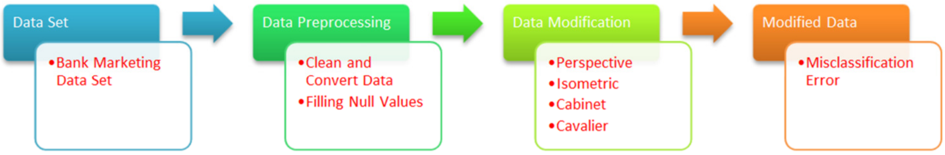

Fig. 1. Workflow of the proposed privacy preservation model.

A. *Perspective Projection*

Perspective projection is used to represent the three-dimensional image of an object on a two-dimensional surface as it is visible to the naked human eye. It is also known as the perspective view. In this projection, lines do not remain parallel and converge at a point called the center of projection. The main advantage of this method is that it achieves privacy with a minimal error rate. In this projection, the projected point is obtained by converging line points of intersection with the plane of the screen.

B. Cavalier Projection

An oblique projection technique called cavalier projection is used to represent three-dimensional objects in two dimensions. It is beneficial for depicting objects with flat faces, such as buildings or furniture, because it strikes the right balance between realism and simplicity. Unlike orthographic projections (such as isometric or dimetric projections), cavalier projections employ angled projection lines rather than perpendicular ones. With cavalier Projection, all features parallel to the projection plane are displayed with their true lengths, and projection lines are angled at a 45^o angle to the plane of projection. This distorts receding dimensions (depth), making the drawing more straightforward to understand than complex projections.

C. Isometric Projection

To depict three-dimensional objects in two dimensions, isometric projection is frequently used in video games, architectural drawings, and engineering diagrams. It is an axonometric projection in which the object is rotated around its axes to show all three dimensions in one view. Isometric projection, in contrast to perspective projection, which reproduces the appearance of objects as they would appear to the human eye, preserves equal scale along each axis, producing parallel lines that stay parallel and consistent angles throughout the image. When an object is projected in an isometric manner, it is rotated around two of its axes relative to the viewer by predetermined angles, typically 120^o. This provides a sense of depth without distortion. Isometric projection uses mathematical transformations to project three-dimensional coordinates onto a two-dimensional plane. This enables the accurate representation of objects in a visually appealing and understandable format.

D. Cabinet Projection

Cabinet projection is a type of axonometric projection that represents three-dimensional objects in two dimensions. The angle at which the object is rotated in relation to the viewer is different from that of the isometric projection despite their similarities. Cabinet projection rotates the object by a specific angle (typically 45^o) about one axis relative to the viewer while the other two axes remain perpendicular to the viewing plane. Because of this, one axis appears shorter than it actually is, but the other two axes are scaled correctly. Therefore, when compared to the isometric projection, cabinet projection produces a more realistic sense of depth. To achieve the desired effect, cabinet projection requires different rotation angles than isometric projection, but it still requires similar mathematical transformations. This makes it possible to accurately depict objects while maintaining an eye-catching sense of depth and proportion.

III. EXPERIMENTAL ANALYSIS

The Bank marketing dataset is used to verify the accuracy and efficiency of the proposed projection methods. The bank marketing dataset contains the details of 41188 phone calls used for direct marketing campaigns of a Portuguese banking institution. It has 16 input attributes, seven of which are numerical and nine of which are categorical. The output variable is a binary class attribute with values yes or no. Only numeric attributes are considered for perturbation. Figure 2 and Table III summarize the output of the proposed projection-based approaches. The misclassification error rate of the suggested approaches is shown in Figure 3 and Table IV.

TABLE III. COMPARISON OF ORIGINAL DATA WITH PERTURBED DATA

Original data	**Projection Techniques**			
	Perspective	**Isometric**	**Cabinet**	**Cavalier**
57	77	57	72	104
56	58	40	74	79
45	47	50	57	83
37	50	48	52	56
25	30	28	31	45

TABLE IV. MISCLASSIFICATION ERROR FOR DIFFERENT CLUSTER SIZES

Projection techniques	**Misclassification Error**			
	K=1	**K=2**	**K=3**	**K=4**
Perspective	0.02	0.04	0.03	0.03
Isometric	0.05	0.08	0.06	0.06
Cabinet	0.07	0.06	0.05	0.08
Cavalier	0.05	0.06	0.06	0.07

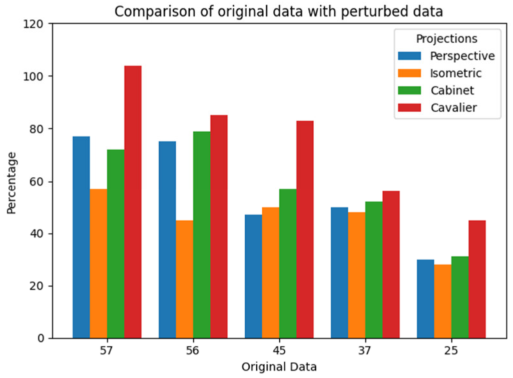

Fig. 2. Output comparison.

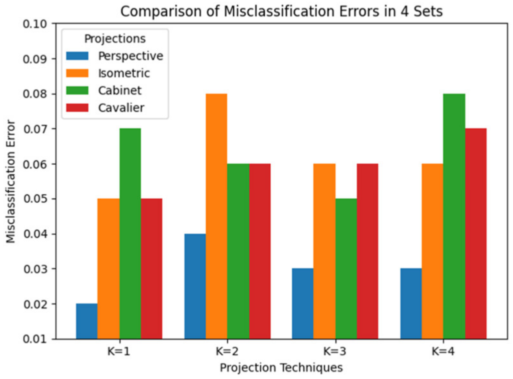

Fig. 3. Misclassification error rate.

Several recent studies from the literature were considered for this study for comparison. Figure 4 and Table V show the comparison result of the considered studies with the proposed model. The outcomes highlight the competitive performance of the proposed method.

TABLE V. PERFORMANCE COMPARISON WITH EXISTING TECHNIQUES

Ref.	Methodology	Accuracy
[1]	Fuzzy logic	89
[2]	Normalization	92
[3]	Geometrical transformation	94.6
[5]	Substitution	95.5
[7]	Artificial neural networks	96
Proposed	**Projection-Based Data Transformation**	**98**

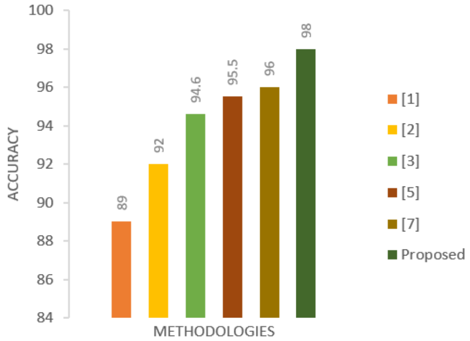

Fig. 4. Accuracy comparison.

V. CONCLUSION

This paper proposes various projection-based privacy-preserving techniques. Many data mining models can be used to test these strategies. Preserving privacy appears to be a crucial aspect of data mining, as it restricts the release of sensitive information before using data processing techniques. The transformation matrices are used to alter the data using the suggested methods, such as perspective, cabinet, cavalier, and isometric projections. The efficiency and error rate over the projected approach has been validated using the Bank Marketing data set. Using k-means clustering and the mathematical misclassification error function to calculate error rate, the aforementioned proposed methods were validated. According to the experimental results, the perspective projection method has a lower error rate than the others, demonstrating its high privacy-preserving capability. Future research should focus on creating context-aware privacy models that take into account the sensitivity of information in various contexts and applications, as well as providing users with tools to understand and control how their data is used and shared in data mining processes.

REFERENCES

[1] B. Karthikeyan, G. Manikandan, and V. Vaithiyanathan, "A fuzzy based approach for privacy preserving clustering," *Journal of Theoretical and Applied Information Technology*, vol. 32, no. 2, pp. 118–122, 2011.

[2] G. Manikandan, N. Sairam, S. Sharmili, and S. Venkatakrishnan, "Achieving Privacy in Data Mining Using Normalization," *Indian Journal of Science and Technology*, vol. 6, no. 4, pp. 4268–4272, Apr. 2013, https://doi.org/10.17485/ijst/2013/v6i4.16.

[3] G. Manikandan, N. Sairam, S. Jayashree, and C. Saranya, "Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation," *Middle-East Journal of Scientific Research*, vol. 14, no. 1, pp. 107–111, 2013, https://doi.org/10.5829/idosi.mejsr.2013.14.1.9313.

[4] C. Saranya and G. Manikandan, "A Study on Normalization Techniques for Privacy Preserving Data Mining," *International Journal of Engineering and Technology*, vol. 5, no. 3, pp. 2701–2704, 2013.

[5] G. Manikandan, N. Sairam, V. Harish, and N. Saikumar, "A substitution based approach for ensuring medical data privacy," *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, vol. 7, no. 2, pp. 1136–1139, Jan. 2016.

[6] G. Manikandan, N. Sairam, V. Harish, and N. Saikumar, "Survey on the use of fuzzy membership functions to ensure data privacy," *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, vol. 7, no. 3, pp. 344–348, Jan. 2016.

[7] D. Niranjan, G. Manikandan, N. Sairam, V. Harish, and N. Saikumar, "Ensuring privacy in data mining using neural networks," *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, vol. 7, no. 4, pp. 1262–1267, Jan. 2016.

[8] Y. Xiao and H. Li, "Privacy Preserving Data Publishing for Multiple Sensitive Attributes Based on Security Level," *Information*, vol. 11, no. 3, Mar. 2020, Art. no. 166, https://doi.org/10.3390/info11030166.

[9] P. S. Rao and S. Satyanarayana, "Privacy preserving data publishing based on sensitivity in context of Big Data using Hive," *Journal of Big Data*, vol. 5, no. 1, Jul. 2018, Art. no. 20, https://doi.org/10.1186/s40537-018-0130-y.

[10] J. Song, Q. Zhong, W. Wang, C. Su, Z. Tan, and Y. Liu, "FPDP: Flexible Privacy-Preserving Data Publishing Scheme for Smart Agriculture," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17430–17438, Dec. 2021, https://doi.org/10.1109/JSEN.2020.3017695.

[11] C. S.-H. Eom, C. C. Lee, W. Lee, and C. K. Leung, "Effective privacy preserving data publishing by vectorization," *Information Sciences*, vol. 527, pp. 311–328, Jul. 2020, https://doi.org/10.1016/j.ins.2019.09.035.

[12] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Security and Communication Networks*, vol. 9, no. 16, pp. 3219–3228, 2016, https://doi.org/10.1002/sec.1527.

[13] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase, "KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes," *Information Security Journal: A Global Perspective*, vol. 26, no. 3, pp. 121–135, May 2017, https://doi.org/10.1080/19393555.2017.1319522.

[14] V. S. Naresh and M. Thamarai, "Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, 2023, Art. no. e1490, https://doi.org/10.1002/widm.1490.

[15] S. Madan and P. Goswami, "A Privacy Preserving Scheme for Big data Publishing in the Cloud using k-Anonymization and Hybridized Optimization Algorithm," in *International Conference on Circuits and Systems in Digital Enterprise Technology*, Kottayam, India, Dec. 2018, pp. 1–7, https://doi.org/10.1109/ICCSDET.2018.8821140.

[16] B. B. Mehta and U. P. Rao, "Improved *l*-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1423–1430, Apr. 2022, https://doi.org/10.1016/j.jksuci.2019.08.006.

[17] P. Zhao, J. Li, F. Zeng, F. Xiao, C. Wang, and H. Jiang, "ILLIA: Enabling k -Anonymity-Based Privacy Preserving Against Location Injection Attacks in Continuous LBS Queries," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1033–1042, Apr. 2018, https://doi.org/10.1109/JIOT.2018.2799545.

[18] T. Qamar, N. Z. Bawany, and N. A. Khan, "EDAMS: Efficient Data Anonymization Model Selector for Privacy-Preserving Data Publishing," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5423–5427, Apr. 2020, https://doi.org/10.48084/etasr.3374.

[19] M. Atif, Z. H. Khand, S. Khan, F. Akhtar, and A. Rajput, "Storage Optimization using Adaptive Thresholding Motion Detection,"

Engineering, Technology & Applied Science Research, vol. 11, no. 2, pp. 6869–6872, Apr. 2021, https://doi.org/10.48084/etasr.3951.

[20] M. O. Al-Dwairi, A. Y. Hendi, and Z. A. AlQadi, "An Efficient and Highly Secure Technique to Encrypt and Decrypt Color Images," *Engineering, Technology & Applied Science Research*, vol. 9, no. 3, pp. 4165–4168, Jun. 2019, https://doi.org/10.48084/etasr.2525.

[21] M. Rathi and A. Rajavat, "Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 9s, pp. 351–367, Jul. 2023.

[22] N. Hrovatin, A. Tosic, M. Mrissa, and B. Kavsek, "Privacy-Preserving Data Mining on Blockchain-Based WSNs," *Applied Sciences*, vol. 12, no. 11, Jan. 2022, Art. no. 5646, https://doi.org/10.3390/app12115646.

[23] P. R. S. Moro, "Bank Marketing." UCI Machine Learning Repository, 2014, https://doi.org/10.24432/C5K306.