information

MDPI

Article

Encrypting and Preserving Sensitive Attributes in Customer Churn Data Using Novel Dragonfly Based Pseudonymizer Approach

Kalyan Nagaraj *, Sharvani GS and Amulyashree Sridhar

Department of Computer Science and Engineering, RV College of Engineering, Bangalore 560059, India
* Correspondence: kalyan1991n@gmail.com

Received: 27 July 2019; Accepted: 28 August 2019; Published: 31 August 2019

Abstract: With miscellaneous information accessible in public depositories, consumer data is the knowledgebase for anticipating client preferences. For instance, subscriber details are inspected in telecommunication sector to ascertain growth, customer engagement and imminent opportunity for advancement of services. Amongst such parameters, churn rate is substantial to scrutinize migrating consumers. However, predicting churn is often accustomed with prevalent risk of invading sensitive information from subscribers. Henceforth, it is worth safeguarding subtle details prior to customer-churn assessment. A dual approach is adopted based on dragonfly and pseudonymizer algorithms to secure lucidity of customer data. This twofold approach ensures sensitive attributes are protected prior to churn analysis. Exactitude of this method is investigated by comparing performances of conventional privacy preserving models against the current model. Furthermore, churn detection is substantiated prior and post data preservation for detecting information loss. It was found that the privacy based feature selection method secured sensitive attributes effectively as compared to traditional approaches. Moreover, information loss estimated prior and post security concealment identified random forest classifier as superlative churn detection model with enhanced accuracy of 94.3% and minimal data forfeiture of 0.32%. Likewise, this approach can be adopted in several domains to shield vulnerable information prior to data modeling.

Keywords: customer data; churn analysis; privacy; ensemble approach; data mining

1. Introduction

Advancement in technology has gathered immense data from several sectors including healthcare, retail, finance and telecommunication. Quantifiable information is captured from consumers to gain valuable insights in each of these sectors [1]. Furthermore, augmented usage of mobile spectrums has paved way for tracing activities and interests of consumers via numerous e-commerce apps [2]. Among multiple sectors tracking consumer data, telecommunication domain is a major arena that has accustomed to several developments ranging from wired to wireless mode, envisioning a digital evolution [3]. Such progressive improvements have generated data in video and voice formats, making massive quantities of customer data accessible to telecom operators. As ratification, Telecom Regulatory Authority of India (TRAI) released a report for the month January 2019 reflecting about 1022.58 million active wireless subscribers in the country [4]. With 5G wireless technologies being appraised as the future generation spectrum, this number is expected to further upsurge [5]. Extracting useful patterns from such colossal data is the ultimate motive of telecom service providers, for understanding customer behavioral trends. Likewise, these patterns also aids in designing personalized services for targeted patrons based on their preceding choices [6]. These preferences further upscale the revenue of a

product by identifying high-value clients. Another substantial parameter accessed by telecom players is churn rate. Churn rate often impacts the marketability of a telecom product.

Churn rate ascertains the extent of subscribers a telecom operator loses to its competitors in a timely manner [7]. Despite gaining new customers, telecom providers are suffering due to churn loss, as it is a well-known fact that retaining old consumers is an easier task than attracting new ones [8]. Henceforth, churn prediction and customer retention are preliminary requirements for handling churn [9]. Over the years, several methods have been devised for detecting customer churn. However, predictions accomplished through mathematical models have gained superior performance in identifying churn [10]. These models are devised by examining the behavior of customer attributes towards assessment of churn. Even though these techniques have been widely accepted for predicting churn rate, there is a downside associated with such predictions.

Analyzing churn features from consumer data invades sensitive information from subscribers. Consumer privacy is sacrificed to decipher the 'value' within data to improve digital marketing and in turn the revenue [11]. Furthermore, such data may be shared with third party vendors for recognizing the ever growing interests of consumers. Despite the existence of TRAI recommendations on data protection for voice and SMS services, it is difficult to implement them in real-time, as data is maintained concurrently with telecom operators and subscribers [12]. Hence, it is advisable to capture only relevant personal details with the consent of consumers prior to data analysis [13]. The same principle can be applied to preserve information prior to churn analysis. This twofold technique ensures data is secured at primal level prior to domain analysis.

Before preserving sensitive information, it is important to scrutinize the attributes which are elusive in the dataset. In this direction, feature selection techniques are being adopted for identifying these attributes. Feature selection identifies the subset of features which are relevant and independent of other attributes in the data [14]. Once sensitive attributes have been identified after feature selection, it is crucial to preserve this information prior to data modeling. In this context, privacy preserving data mining (PPDM) techniques have gathered immense popularity over the years for their proficiencies to secure data. PPDM approaches preserve data integrity by converting the vulnerable features into an intermediate format which is not discernible by malicious users [15,16]. Some of the popular PPDM techniques implemented for ensuring privacy include k-anonymity, l-diversity, t-closeness, ε-differential privacy and personalized privacy. Each of these techniques employs different phenomenon to preserve vulnerable information. Even though k-anonymization is an effective approach it ignores sensitive attributes, l-diversity considers sensitive attributes yet distribution of these features are often ignored by the algorithm [17]. Coming to t-closeness, correlation among identifiers decreases as t-value increases due to distribution of attributes. ε-differential and personalized privacy are difficult to implement in real time with dependency on the user for knowledgebase. Hence, there is a need to devise an enhanced PPDM model which is capable of preserving vulnerable information without compromising on data quality, privacy measures and simplicity of the approach [18].

In this direction, the current study adopts a twofold approach for securing sensitive information prior to churn prediction. Initially, dragonfly algorithm is applied for optimizing profound attributes from churn data. Furthermore, privacy preserving is employed on the identified features using pseudonymization approach for avoiding privacy breach. Consequently, the performance of this novel twofold approach is compared with conventional methods to determine information camouflage. Once sensitive details are secured from the churn dataset, data mining models are employed to detect the occurrence of churn among consumers. These models are furthermore analyzed to discover information loss prior and post the ensemble approach. Such a twofold technique ensures data privacy is obscured without disclosing the context.

2. Literature Survey

2.1. Churn Detection via Mathematical Modeling

As mentioned previously, churn rate estimation is crucial for telecom providers to access the customer satisfaction index. Churn assessment via mathematical models has been adopted extensively over the years to predict customer migration. Numerous studies have estimated churn rate using linear and non-linear mathematical models. Some of the significant studies are discussed in Table 1.

Despite the existence of such outstanding classifiers, there is a need to scrutinize these models to ensure that no susceptible information is leaked. Thereby, the next section elaborates on importance of privacy preserving techniques towards information conservation.

2.2. PPDM Approaches towards Information Security

This section introduces the implications of several PPDM approaches towards masquerade of sensitive data. Several PPDM techniques have been designed to mitigate security attacks on data including anonymization, randomization and perturbation [18]. Some of the predominant variants of anonymization (i.e., k-anonymity, l-diversity, t-closeness) and other techniques are discussed in Table 2.

Extensive literature survey in churn prediction and privacy preservation has discovered the possibility for employing PPDM techniques towards securing sensitive churn attributes. This two fold approach of combining information security in churn detection is unique in preserving sensitive data attributes prior to modeling.

Table 1. List of substantial churn detection studies.

Studies Performed	Mathematical Model/s Adopted	Substantial Outcomes	Limitations/Future Scope
[19]	Linear Regression	The study achieved 95% confidence interval in detecting customer retention	-
[20]	Legit regression, Boosting, Decision trees, neural network	The non-linear neural network estimated better customer dissatisfaction compared to other classifiers	-
[21]	Support Vector Machine (SVM), neural network, legit regression	SVM outperformed other classifiers in detecting customer churn in online insurance domain	Optimization of kernel parameters of SVM could further uplift the predictive performance
[22]	Random forest, Regression forest, logistic & linear regression	Results indicated that random and regression forest models outperformed with better fit compared to linear techniques	-
[23]	AdaBoost, SVM	Three variants of AdaBoost classifier (Real, Gentle, Meta) predicted better churn customers from credit debt database compared to SVM	-
[24]	SVM, artificial neural network, naïve bayes, logistic regression, decision tree	SVM performed enhanced customer churn detection compared to other classifiers	-
[25]	Improved balanced random forest (IBRF), decision trees, neural networks, class-weighted core SVM (CWC-SVM)	IBRF performed better churn prediction on real bank dataset compared to other classifiers	-
[26]	Partial least square (PLS) classifier	The model outperforms traditional classifiers in determining key attributes and churn customers	-
[27]	Genetic programming, Adaboost	Genetic program based Adaboosting evaluates churn customers with better area under curve metric of 0.89	-
[28]	Random forest, Particle Swarm Optimization (PSO)	PSO is used to remove data imbalance while random forest is implemented to detect churn on reduced dataset. The model results in enhanced churn prediction	-
[29]	Naïve Bayes, Bayesian network & C4.5	Feature selection implemented by naïve Bayes and Bayesian network resulted in improved customer churn prediction	Overfitting of minor class instances may result in false predictions. Hence balancing the data is to be performed extensively.
[30]	Decision tree, K-nearest neighbor, artificial neural network, SVM	Hybrid model generated from the classifiers resulted in 95% accuracy for detecting churn from Iran mobile company data	-
[31]	Rough set theory	Rules are extracted for churn detection. Rough set based genetic algorithm predicted churn with higher efficacy	-

Table 1. *Cont.*

Studies Performed	Mathematical Model/s Adopted	Substantial Outcomes	Limitations/Future Scope
[32]	Bagging, boosting	Among multiple classifiers compared for churn detection, bagging and boosting ensembles performed better prediction	-
[33]	Dynamic behavior models	Spatio-temporal financial behavioral patterns are known to influence churn behavior	Possibility of bias in the data may affect the predictive performances
[34]	Naïve Bayes	Customer churn prediction is detected from publically available datasets using naïve Bayes classifier	The current approach can be implemented for detecting bias and outliers effectively
[35]	Decision tree, Gradient boosted machine tree, Extreme gradient boost and random forest	The extreme gradient boosting model resulted in area under curve (AUC) value of 89% indicating an enhanced churn classification rate compared to other classifiers	-

Table 2. Significance of PPDM approaches towards data security.

PPDM Techniques	Features of the Technique	Studies Adopted Based on the Technique
k-anonymity	The technique masks the data by suppressing the vulnerable instances and generalizing them similar to other (k-1) records.	[36–44]
l-diversity	Extension of k-anonymity technique which reduces granularity of sensitive information to uphold privacy	[17,45–48]
t-closeness	Extension of l-diversity that reduces granularity by considering distribution of sensitive data	[49–53]
Randomization	Randomizes the data instances based on its properties to result in distorted data aggregates	[54–57]
Perturbation	Noise is added to data to ensure that sensitive information is not disclosed	[58–61]
Pseudonymization	Reversible pseudonyms replaces sensitive information in data to avoid data theft	[62–64]

3. Materials and Methods

This section highlights the methodology adopted for churn detection accompanied by security concealment of susceptible information.

3.1. Churn Data Collection

From massive population of customer indicators in telecom sector, a suitable sample is to be identified to analyze the vulnerable patterns. For this reason, publically available churn dataset is selected from web repository Kaggle. This dataset is designed to detect customer churn based on three classes of attributes. The first class identifies the services a customer is accustomed to, second class of attributes identifies the information of financial status towards bill payment and third class discusses about the demographic indications for a customer. Spanned across these three classes are 20 data features for predicting churn behavior. This dataset is downloaded in .csv file format for further processing [65].

3.2. Initial Model Development

Mathematical models are developed to predict influence of data attributes in churn analysis. Conventional algorithms like logistic regression, Support Vector Machine (SVM), naïve Bayes, bagging, boosting and random forest classifiers are employed on the data using modules available in R programming language. Cross validation technique is adopted in 10-fold manner for validating the outcomes from every learner.

3.3. Assessment of Performance of Models

Of the learners implemented in previous step, the best performing model is ascertained using certain statistical parameters including True Positive Rate (TPR), accuracy, F-measure and Root Mean Square Error (RMSE). The formulae's of these metrics are defined below:

True positive is the metric where a mathematical model predicts the churn instances accurately while false negative reflects the prediction of churn cases incorrectly as non-churn by a model. TPR is the ratio of true positive churn instances in the dataset to the summation of true positive and false negative churn cases.

$$TPR = \frac{True\ Postive\ instances\ of\ customer\ churn}{True\ positive\ instances\ of\ churn + False\ negative\ instances\ of\ churn} \quad (1)$$

Accuracy is the ratio of predictions of churn and non-churn instances from a mathematical model against the total churn instances in the dataset. Higher the value better is the performance of a classifier.

$$Accuracy = \frac{Total\ instamces\ of\ churn\ and\ non-churn\ predictions\ by\ models}{Total\ instances\ of\ churn\ and\ non-churn\ in\ dataset} \quad (2)$$

Root mean square error (RMSE) is the standard deviation of predicted churn customers by mathematical model compared to actual churn customers in the dataset. The metric ranges from 0 to 1. Lesser the value better is the performance of a model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted\ churn\ instances\ by\ models - Actual\ churn\ instances\ in\ dataset)^2}{Total\ churn\ an\ d\ non-churn\ instances\ in\ dataset}} \quad (3)$$

F-measure is defined as the harmonic mean of recall and precision. Recall indicates the score of relevant churn customers identified from churn dataset, while precision specifies the score of identified customers who are turned out to be churned.

$$F - measure = 2 \times \left(\frac{recall \times precision}{recall + precision} \right) \tag{4}$$

3.4. Preserving Sensitive Churn Attributes Using Dragonfly Based Pseudonymization

Applying mathematical models as in previous steps without securing the vulnerable attributes will result in an intrinsic security threat. Hence, a twofold technique is adopted to withhold the privacy, prior to churn prediction. This dual procedure involves optimal feature selection followed by privacy preservation.

Feature selection plays a significant role in this study owing to two fundamental motives. Primarily, a data mining model cannot be designed accurately for a dataset having as many as 20 features, as there is high probability for dependency among attributes [66]. Another vital intention for adopting feature selection is to identify vulnerable attributes (if any) prior to data modeling.

In this context, dragonfly algorithm is employed on the dataset for attribute selection. This algorithm is selected amongst others because of its preeminence in solving optimization problems, specifically feature selection [67]. Dragonfly algorithm is conceptualized based on dual behavioral mode of dragonfly insects. They remain either in static (in case of hunting for prey) or dynamic (in case of migration) mode. This indistinguishable behavior of dragonflies in prey exploitation and predator avoidance is modeled using five constraints based on the properties of separation, alignment, cohesion, food attraction and enemy distraction. These parameters together indicate the behavior of a dragonfly. The positions of these flies in search space are updated using two vectors, position vector (Q) and step vector (ΔQ) respectively [68].

All these parameters are represented mathematically as follows:

a. *Separation* (S_i): It is defined as the difference between current position of an individual dragonfly (Z) and ith position of the neighboring individual (Z_i) summated across total number of neighbors (K) of a dragonfly;

$$S_i = -\sum_{i=1}^{K} Z - Z_i \tag{5}$$

b. *Alignment* (A_i): It is defined as the sum total of neighbor's velocities (V_k) with reference to all the neighbors (K);

$$A_i = \frac{\sum_{i=1}^{K} V_k}{K} \tag{6}$$

c. *Cohesion* (C_i): It is demarcated as the ratio of sum total of neighbor's ith position (Z_i) of a dragonfly to all the neighbors (K), which is subtracted from the current position of an individual (Z) fly;

$$C_i = \left(\frac{\sum_{i=1}^{K} Z_i}{K} \right) - Z \tag{7}$$

d. *Attraction towards prey/food source* (P_i): It is the distance calculated as the difference between current position of individual (Z) and position of prey (Z^+);

$$P_i = Z^+ - Z \tag{8}$$

e. *Avoidance against enemy* (E_i): It is the distance calculated as the difference between current position of individual (Z) and position of enemy (Z^-).

$$E_i = Z^- - Z \tag{9}$$

The step vector (ΔQ) is further calculated by summating these five parameters as a product of their individual weights. This metric is used to define the locus of dragonflies in search space across iterations denoted by:

$$\Delta Q_{t+1} = (aA_i + sS_i + cC_i + pP_i + eE_i) + w\Delta Q_t \tag{10}$$

Here A_i, S_i, C_i, P_i and E_i denote the alignment, separation, cohesion, prey and enemy coefficients for ith individual dragonfly. Correspondingly a, s, c, p, e denotes the weights from alignment, separation, cohesion, prey and enemy parameters for ith individual. These parameters balance the behavior of prey attraction and predator avoidance. Here, w represents the inertia weight and t represents number of iterations. As the algorithm recapitulates, converge is assured when the weights of the dragonfly constraints are altered adaptively. Concurrently, position vector is derived from step vector as follows:

$$Q_{t+1} = Q_t + \Delta Q_{t+1} \tag{11}$$

Here, t is defined to denote the present iteration. Based on these changes, the dragonflies optimize their flying paths as well. Furthermore, dragonflies adapt random walk (Lèvy flight) behavior to fly around the search space to ensure randomness, stochastic behavior and optimum exploration. In this condition, the positions of dragonflies are updated as mentioned below:

$$Q_{t+1} = Q_t + Lievy(d) \times Q_t \tag{12}$$

Here d denotes the dimension for position vector of dragonflies.

$$Lievy(q) = 0.01 \times \frac{m_1 \times \sigma}{|m_2|^{\frac{1}{\beta}}} \tag{13}$$

The Lèvy flight, *Lèvy* (q) defines two parameters m_1 and m_2 denoting two random number in the range [0, 1], while β is a constant. Furthermore, σ is calculated as follows:

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin\frac{\pi\beta}{2}}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}} \right)^{\frac{1}{\beta}} \tag{14}$$

Here, $\Gamma(x) = (x-1)!$

Based on these estimates, the dragonfly algorithm is implemented on the churn dataset. Prior to feature selection, the dataset is partitioned into training (70%) and test (30%) sets. Training data is used initially for feature selection, while test data is applied on results derived from feature selection to validate the correctness of the procedure adopted. Dragonfly algorithm comprises of the following steps:

(i) *Initialization of parameters*: Initially, all the five basic parameters are defined randomly to update the positions of dragonflies in search space. Each position corresponds to one feature in the churn dataset which needs to be optimized iteratively.

(ii) *Deriving the fitness function*: After parameters are initialized, the positions of the dragonflies are updated based on a pre-defined fitness function F. The fitness function is defined in this study based on objective criteria. The objective ensures the features are minimized iteratively without compromising on the predictive capabilities of selected features towards churn detection. This objective is considered in the defined fitness function adopted from [68], using weight factor,

w_j to ensure that maximum predictive capability is maintained after feature selection. The fitness function F is defined below. Here, Pred represents the predictive capabilities of the data features, w_j is the weight factor which ranges between [0, 1], L_a is the length of attributes selected in the feature space while L_s is the sum of all the churn data attributes;

$$F = max\left(Pred + w_j\left(1 - \frac{L_a}{L_s}\right)\right) \tag{15}$$

(iii) *Conditions for termination*: Once the fitness function F fails to update neighbor's parameters for an individual dragonfly, the algorithm terminates on reaching the best feature space. Suppose the best feature space is not found, the algorithm terminates on reaching the maximum limit for iterations.

The soundness of features derived from dragonfly algorithm is further validated using a wrapper-based random forest classifier, Boruta algorithm available in R programming language. The algorithm shuffles independent attributes in the dataset to ensure correlated instances are separated. It is followed by building a random forest model for merged data consisting of original attributes. Comparisons are further done to ensure that variables having higher importance score are selected as significant attributes from data [69]. Boruta algorithm is selected because of its ability in choosing uncorrelated pertinent data features. Henceforth this confirmatory approach ensures relevant churn attributes are selected prior to privacy analysis.

From the attributes recognized in previous step, vulnerable features needs to be preserved prior to churn analysis. For this reason, privacy preserving algorithms are to be tested. These algorithms must ensure that preserved attributes are not re-disclosed after protection (i.e., no re-disclosure) and must also ensure that original instance of data is recoverable at the time of need (i.e., re-availability). Techniques like k-anonymization prevent re-disclosure, however if reiterated, information can be recovered leading to risk of privacy breach. Hence such techniques cannot be adopted while securing sensitive information from churn data. Thereby, enhanced algorithms of privacy preservation are to be adopted for supporting re-availability and avoiding re-disclosure of the data. One such privacy mechanism to conserve data is by pseudonymization [70]. Pseudonymization is accomplished by replacing vulnerable attributes with consistent reversible data such that information is made re-available and not re-disclosed after replacement. This technique would be suitable for protecting churn attributes.

Pseudonymization is applied by replacing vulnerable churn identifiers pseudonyms or aliases so that the original customer data is converted to an intermediate reversible format. These pseudonyms are assigned randomly and uniquely to each customer instance, by avoiding duplicate cases. Pseudonymization is performed in R programming language for features derived from dragonfly algorithm. Furthermore, to avoid attacks due to knowledge of original data (i.e., background knowledge attack) the pseudonyms are encrypted using unique hash indices. Suppose the pseudonyms need to be swapped back to original data decryption is performed. Thereby encryption enabled pseudonymization ensures that data breach is prohibited from background knowledge attacks [71]. Additionally, performance of the pseudonymization approach is compared with other privacy preserving models to access efficiency of data preservation.

3.5. Development of Models Post Privacy Analysis

Once the sensitive information is secured using the integrated dragonfly based pseudonymization approach, mathematical models are once again developed for detecting churn cases. Previously employed conventional algorithms are applied once again on the preserved churn dataset. The churn rate is analyzed from these models for the unpreserved relevant churn attributes as performed previously.

3.6. Detection of Information Loss Prior and Post Privacy Preservation

It is important to ensure that any privacy preservation approach is not associated with loss of appropriate information from original data. For this reason, performance of the churn prediction algorithms are accessed recursively using statistical techniques to ensure that there is no significant statistical difference in the prediction abilities of the classifiers before and after privacy preservation. For this reason, the research hypothesis is framed with the following assumptions:

Null Hypothesis (H_0). *There is no statistical difference in the performance of the churn detection models prior and post privacy analysis.*

Alternative Hypothesis (H_1). *There is substantial statistical difference in the performance of the churn detection models prior and post privacy analysis.*

Prior to research formulation, it is assumed that H_0 is valid. Furthermore this convention is validated by performing Student's *t*-test between the pre and post-security development models.

4. Results

This section highlights the significant outcomes derived from churn detection from telecom customer's dataset.

4.1. Processing of Churn Dataset

The customer churn data downloaded from Kaggle is analyzed to identify instances of customer churn vs. non-churn. Three classes of attributes are responsible for detecting customer churn based on accustomed services (i.e., internet, phone, streaming content access, online support and device protection) account details (i.e., payment mode, monthly bills, contract mode, paperless bills and total billing) and demographic details of a customer (i.e., age, gender, dependents and partners). These classes comprises of 20 attributes in the dataset which collectively predict customer churn. The dataset all together includes 7043 instances having 1869 occurrences of customer churn, while remaining are loyal customers. The churn prediction approach adopted in this work is shown in Figure 1. Furthermore, the distribution of churn attributes in the dataset is shown in Figure 2.

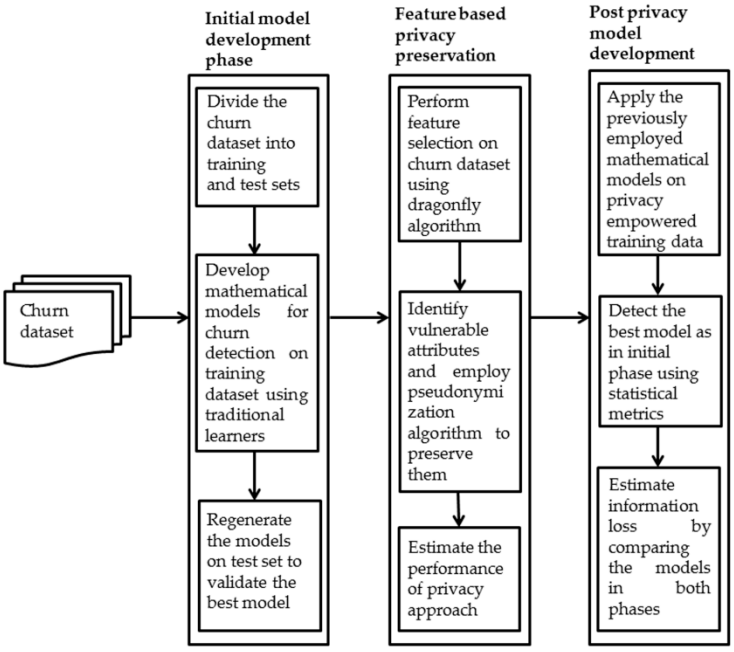

Figure 1. Overview of privacy imposed churn detection approach.

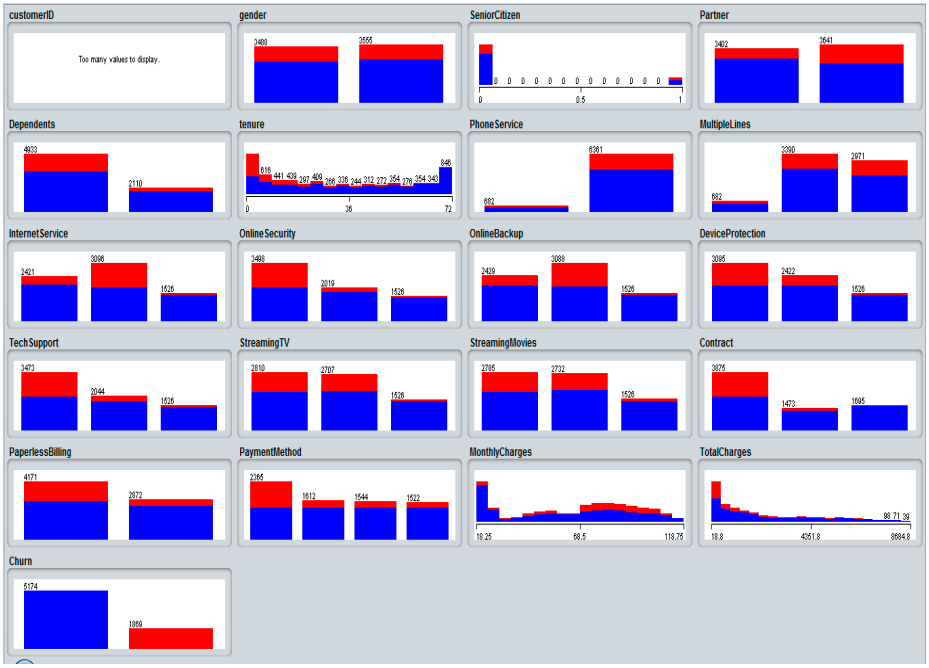

Figure 2. The distribution of attributes in churn dataset; Here red indicates churn customers while blue indicates non-churn customers. The figure is derived from Weka software [72].

4.2. Initial Model Development Phase

Once data is collected, churn prediction is performed initially on all the data attributes. Several data mining models including logistic regression, naïve Bayes, SVM, bagging, boosting and random forest classifiers are adopted for churn detection. Each of these models is developed in R programming language using dependencies available in the platform. Statistical parameters like accuracy and F-measure suggested that random forest classifier outperformed other classifiers in terms of churn detection. The performance of these classifiers is shown in Table 3.

Table 3. The performance of models in initial phase of churn detection.

Sl. No	Classifier	R Language Dependency	True Positive Rate	Accuracy	RMSE	F-Measure
1.	Logistic Regression	glm	0.887	0.793	0.398	0.793
2.	Naïve Bayes	naivebayes	0.893	0.791	0.334	0.801
3.	SVM	e1071	0.910	0.839	0.298	0.835
4.	Bagging	adabag	0.912	0.860	0.263	0.866
5.	Boosting	adabag	0.934	0.889	0.191	0.905
6.	Random Forest	randomForest	0.997	0.956	0.112	0.956

However, these models exposed certain sensitive information while detecting customer churn including demographic particulars. Hence, feature selection is to be performed to identify the susceptible churn attributes which needs to be conserved prior to data modeling.

4.3. Feature Selection and Attribute Preservation Using Dragonfly Based Pseudonymizer

To identify and preserve subtle churn attributes, a dual approach is adopted using feature selection and privacy preservation techniques. Feature selection is implemented using dragonfly algorithm by subjecting the data attributes randomly as initial population using 'metaheuristicOpt' dependency in R programming language [73]. The fitness function F is evaluated such that feature set is minimized by

retaining superlative attributes. Hence, the algorithm performs optimization by feature minimization. In this case, schewefel's function is used for defining the objective.

The algorithm works by recognizing the best feature as food source and worst feature as enemy for a dragonfly. Furthermore the neighboring dragonflies are evaluated from each dragonfly based on the five parameters i.e., S_i, A_i, C_i, P_i and E_i respectively. These neighbors are updated iteratively based on radius parameter which increases in linear fashion. The weights updated in turn helps in updating the position of remaining dragonflies. This procedure is iterated until germane churn features are identified from 500 iterations in random orientation. The threshold limit is set to 0.75 (i.e., features above 75% significance w.r.t churn rate are selected) to ignore less pertinent attributes.

Results revealed eight features as significant in churn detection based on the estimated values of fitness function F. The data features ranked as per F is shown in Table 4. The significant attributes identified from the table include contract, total charges, tenure, tech support, monthly charges, online backup, online security, and internet service respectively.

Additionally, these features are also affirmed by wrapper embedded learner, Boruta based on their importance scores in R programming language. The algorithm is iterated 732 times to derive eight significant features based on their importance scores. These attributes are visualized as a plot in Figure 3 revealing their importance scores. Higher the importance score, superior is the impact of a data feature towards churn rate. The plot illustrates features equivalent to dragonfly computation by suggesting alike features as significant churn attributes. However the order of their importance varies across both the computations.

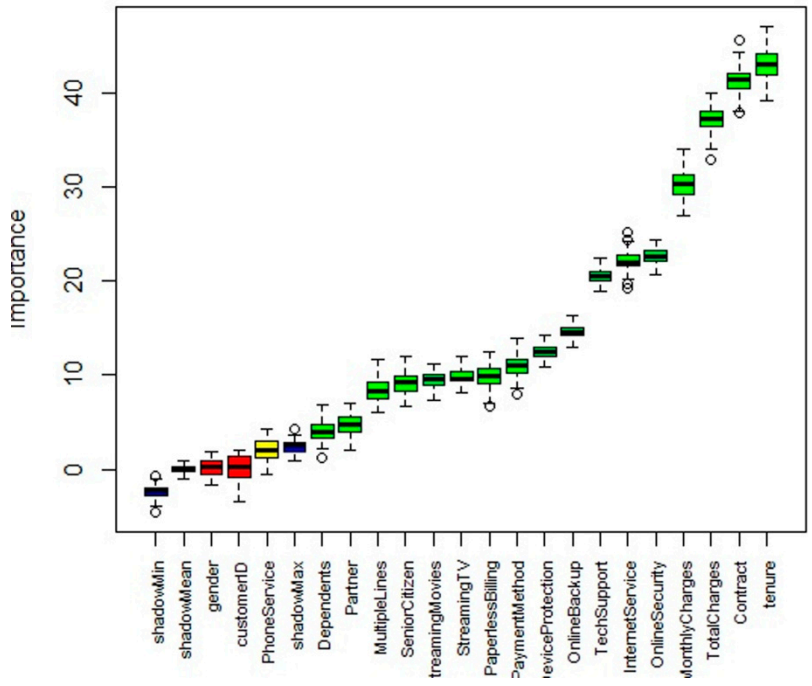

Figure 3. The distribution of churn features based on importance score in Boruta algorithm.

Table 4. Churn features arranged as per their relevance via fitness function.

Sl. No	Feature Name	Feature Description	Feature Category	Fitness Value
1.	Contract	Denotes the contract period of the customer if it is monthly, yearly or for two years	Account information	0.9356
2.	Tenure	Indicates the number of months a customer is patron to the service provider	Account information	0.9174
3.	Total charges	Indicates the total charges to be paid by the customer	Account information	0.9043
4.	Monthly charges	Indicates the monthly charges to be paid by the customer	Account information	0.8859
5.	Tech support	Indicates if the customer has technical support or not, based on the internet service accustomed	Customer services	0.8533
6.	Online security	Indicates if the customer has online security or not, based on the internet service accustomed	Customer services	0.8476
7.	Internet service	Indicates the internet service provider of the customer, which can be either fiber optic, DSL or none of these	Customer services	0.7971
8.	Online backup	Indicates if the customer has online backup or not, based on the internet service accustomed	Customer services	0.8044
9.	Payment method	Denotes the type of payment method. It can be automatic bank transfer mode, automatic credit card mode, electronic check or mailed check	Account information	0.7433
10.	Streaming TV	Denotes if the customer has the service for streaming television or not	Customer services	0.7239
11.	Paperless billing	Denotes if the customer has the service for paperless billing or not	Account information	0.7009
12.	Streaming movies	Indicates if the customer has the service for streaming movies or not	Customer services	0.6955
13.	Multiple lines	Indicates if the customer has service for multiple lines or not	Customer services	0.5487
14.	Senior Citizen	Indicates if the customer is a senior citizen or not	Demographic details	0.5321
15.	Partner	Denotes whether the customer has a partner or not	Demographic details	0.5093
16.	Phone Service	Indicates if the customer has services of the phone or not	Customer services	0.5005
17.	Dependents	Indicates if the customer has any dependents or not	Demographic details	0.4799
18.	Device protection	Denotes if the customer has protection or not for the device	Customer services	0.4588
19.	Gender	Indicates if the customer is male or female	Demographic details	0.3566
20.	Customer ID	A unique identifier given to each customer	Demographic details	0.2967

Of these eight distinctive churn attributes, characteristics like tenure, contract, monthly charges and total charges reveal sensitive information about churn customers. Hence, these attributes needs to be preserved before model development. However, preservation must also ensure that the sensitive information is regenerated when required. For this purpose, pseudonymization technique is employed on the selected attributes using 'synergetr' package in R programming language [74]. Pseudonymization is a privacy preservation technique which replaces the vulnerable information with a pseudonym. The pseudonym (J_i) is an identifier that prevents disclosure of data attributes. Pseudonyms are defined for each vulnerable attribute V_1, V_2, V_3 and V_4 in the dataset, so that original data is preserved. Furthermore these pseudonyms are encrypted using hash indices to avoid privacy loss due to background knowledge attacks. To derive original data from the encrypted pseudonym, decryption key is formulated.

This twofold approach of feature selection and privacy preservation is named as "Dragonfly based pseudonymization". The algorithm is iterated over finite bound of 10,000 random iterations to infer the concealment of churn information. However, if there is no vulnerable attribute present in the data at that instance, the algorithm iterates until its maximum limit and terminates.

The ensemble algorithm adopted in this approach is shown in Algorithm 1.

Algorithm 1 Dragonfly based pseudonymizer

1. Define the initial values of dragonfly population (P) for churn data denoting the boundary limits for maximum number of iterations (n)
2. Define the position of dragonflies (Y_i such that i = 1, 2, ..., n)
3. Define the step vector ΔY_i such that i = 1, 2, ..., n
4. **while** termination condition is not reached **do**
5. Calculate the fitness function F, for every position of dragonfly
6. Identify the food source and enemy
7. Update the values of w, s, a, c, p and e
8. Calculate S_i, A_i, C_i, P_i and E_i using Equations (5)–(9)
9. Update the status of neighboring dragonflies
10. **if** atleast one neighboring dragonfly exists
11. Update the step vector using (10)
12. Update the velocity vector using (11)
13. **else**
14. Update the position vector using (12)
15. Discover if new positions computed satisfy boundary conditions to bring back dragonflies
16. Generate best optimized solution O
17. Input the solution O to pseudonymizer function
18. Define the length of the pseudonym J_i for each vulnerable attribute V_a such that a = 1, 2, ..., n
19. Eliminate duplicate pseudonyms
20. Encrypt with relevant pseudonyms for all data instances of vulnerable attributes
21. Repeat until pseudonyms are generated for all vulnerable attributes
22. Replace the vulnerable attribute V_a with pseudonyms J_i
23. Reiterate until all sensitive information in V_a is preserved
24. Produce the final preserved data instances for churn prediction
25. Decrypt the pseudonyms by removing the aliases to view original churn dataset

4.4. Performance Analysis of Dual Approach

The performance of privacy endured twofold model is to be estimated by comparison with other PPDM models. For this purpose, susceptible features selected from dragonfly algorithm are subjected to privacy preservation by employing randomization, anonymization and perturbation algorithms in R programming platform. The performance of all the models towards feature conservancy is enlisted in Table 5.

Table 5. Comparison of privacy preservation by pseudonymization and other PPDM models.

Sl. No	Features to Be Preserved	Iterations	Pseudonymization	Anonymization	Randomization	Perturbation
1.	Tenure	1000	Preserved	Preserved	Preserved	Not preserved
2.	Contract	1000	Preserved	Preserved	Not preserved	Preserved
3.	Monthly charges	1000	Preserved	Not preserved	Not preserved	Preserved
4.	Total charges	1000	Preserved	Not preserved	Preserved	Preserved

As observed from the table, pseudonymization technique performs with better stability over other algorithms by securing all the key attributes over 1000 random iterations.

4.5. Model Re-Development Phase

The churn dataset derived subsequently after privacy enabled approach with eight essential features having four preserved attributes is taken as input for model re-development phase. The algorithms used in initial model development phase are employed at this point for accessing churn rate among customers. These models are developed in R programming language as of previous phase. The results from model development in re-development phases are enlisted in Table 6.

Table 6. Performance of models in re-development phase after privacy preservation.

Sl. No	Classifier	R Language Dependency	True Positive Rate	Accuracy	RMSE	F-Measure
1	Logistic Regression	glm	0.887	0.788	0.398	0.793
2	Naïve Bayes	naivebayes	0.893	0.780	0.334	0.801
3	SVM	e1071	0.910	0.828	0.298	0.835
4	Bagging	adabag	0.912	0.858	0.263	0.866
5	Boosting	adabag	0.934	0.873	0.191	0.905
6	Random Forest	randomForest	0.997	0.943	0.112	0.956

The table displays random forest algorithm as the best performing classifier with enhanced accuracy. Remaining classifiers are similarly arranged based on their performance abilities in detecting churn rate.

4.6. Estimating Differences in Models Based on Hypothesis

The predictive performance of models in the initial phase is compared with the models in the re-development phase to adjudicate the amount of information loss. For this reason, *t*-test is performed on the churn models from initial and re-development phases. The cutoff value alpha is assumed to be 0.05 for estimating the statistical difference among two categories of learners. Initially, the data churn instances are randomly reshuffled to split them into training (70%) and test (30%) datasets. Mathematical models are generated on training data as per previous iterations and F metric is estimated for all the models. Furthermore, the F value is evaluated on test data to avoid bias. From such random churn population, a sample data is extracted with 500 data instances having 200 churn and 300 non-churn customer cases. *F* metric is computed on this sample set and *t*-test score is evaluated for estimating the validity of H_o w.r.t alpha value.

Furthermore, *p*-value is estimated for both categories of models as shown in Table 7. These p-values are compared with alpha value to estimate statistical differences among the learners. As observed from the table, *p*-values are found to be lesser than the alpha value (i.e., 0.05).

Table 7. Performance of models in re-development phase after privacy preservation.

Group 1 (Initial Models)	Group 2 (Re-Development Models)	Sample Data Size	*p*-Value	t-Score
Logistic regression	Logistic regression	500	0.04	188.86
Naïve Bayes	Naïve Bayes	500	0.04	198.67
SVM	SVM	500	0.03	165.23
Bagging	Bagging	500	0.03	154.89
Boosting	Boosting	500	0.02	99.35
Random Forest	Random Forest	500	0.01	66.95

Henceforth, there is a significant difference in the performance of the models between initial and re-development phases. Thereby, the research hypothesis H_0 is found to be invalid. Alternatively H_1 is accepted in this case to ensure there is difference in information between the two categories of models. These differences in the models ensure the presence of information loss.

4.7. Estimating Information Loss Between Initial and Re-Development Models

The previous step ensured that there is inherent information loss in the models. Hence it is noteworthy to estimate the value. Ideally, there must not be any information loss associated before and after information preservation. However, information loss is observed in this case as data attributes are modified. Modification can happen either due to deletion or addition of information, leading to loss of original data. Hence, information loss is defined for absolute values of churn data instances accordingly by ignoring the sign:

$$Information\ Loss\ (\%) = \frac{(M_{ic} - M_{rc})}{Total\ churn\ instances\ in\ the\ dataset} * 100 \quad (16)$$

Here, M_{ic} indicates the churn predicted by initial models and M_{rc} indicates the churn prediction by re-development phase models.

The tabulation of information loss for the models in both phases is reflected in Table 8. As observed from table, minimal information loss is observed from random forest classifier. This result is in par with other computations as the same classifier has achieved better churn detection in both the phases of model development. Hence, minimal information loss is associated with the model that performs better churn detection.

Table 8. Information loss analysis in two phases of model development.

Sl. No	Classifier	Iterations	Total Churn Instances in the Dataset	Data Instances Detecting Churn in Initial Model Development	Data Instances Detecting Churn After Model Re-Development	Information Loss after Dual Approach (%)
1.	Logistic Regression	900	1869	1203	1109	5.02
2.	Naïve Bayes	900	1869	1301	1245	2.99
3.	SVM	900	1869	1432	1397	1.87
4.	Bagging	900	1869	1645	1657	0.64
5.	Boosting	900	1869	1793	1804	0.58
6	Random Forest	900	1869	1823	1829	0.32

5. Discussion

The current study is designed to analyze the impact of privacy preservation in churn rate prediction. To emphasize on its importance, customer churn dataset is initially retrieved. Based on the analysis, four key findings are derived from this study: (i) A twofold approach designed for feature selection and privacy preservation using dragonfly based pseudonymization algorithm; (ii) set of relevant features

which detect customer churn; (iii) set of vulnerable churn attributes which are to be preserved prior to churn detection and (iv) churn detection using mathematical models.

In context of customer churn detection, the features identified in the current study indicate significance of attribute selection techniques prior to model development. Suppose model development phase precedes feature selection, there is no assurance that pertinent features could be utilized to identify churn. There is high likelihood that predictions are amplified incidentally due to repetitive iterations. To eliminate such cases of bias, the study performs model development prior to feature selection followed by model development post feature selection and privacy preservation. This twofold model development phase ensures that churn is detected without any feature dependency. Furthermore, information loss due to data sanctuary is detected by analyzing models developed in both the phases. This twofold approach seems to be appropriate for safeguarding interest of consumers against viable security threats. Even though several protocols are available to preserve information till date, current data mining approach helps in predicting future possibilities of churn instances based on the existing statistics from mathematical models [30].

However the study has two major limitations. The first limitation is due to the contemplation of a solitary dataset for concealment conservation in churn. Outcomes derived from the individual data analysis are often explicit to the information. These outcomes may not be relevant on a different churn dataset. Hence, global analysis of multidimensional churn datasets helps in validating the outcomes derived after privacy preservation. Such dynamic predictive approaches will provide insights into large-scale security loopholes at data level. The second limitation is due to the small set of vulnerable churn features in the dataset. Suppose the sensitive information is increased by 10 or 100-fold it is not guaranteed that the current approach would perform with same efficacy. The algorithm needs to be tested and fine-tuned on datasets with different feature dimensionalities.

Data driven implementation adopted in this study can be utilized in developing a global privacy impenetrable decision support system for telecom operators and subscribers. To accomplish this undertaking, superfluous investigation is required in further stages. As a practical direction, one can consider relevant features in churn detection for optimization of parameters in distributed mode which aids in elimination privacy ambiguities. Henceforth, several nascent security threats would be filtered out prior to data analysis workflow. An analogous approach can be employed by developing multidimensional privacy preserving models on the data features using centralized and distributed connectivity to provide encryption of multiple formats from subtle information.

Author Contributions: The conceptualization of work is done by K.N. and S.G.; methodology, validation and original draft preparation is done by K.N. and A.S.; Reviewing, editing and supervision is done by S.G.

Funding: This research received no external funding.

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

References

1. Diaz, F.; Gamon, M.; Hofman, J.M.; Kıcıman, E.; Rothschild, D. Online and Social Media Data as an Imperfect Continuous Panel Survey. *PLoS ONE* **2016**, *11*, e014506. [CrossRef] [PubMed]
2. Tomlinson, M.; Solomon, W.; Singh, Y.; Doherty, T.; Chopra, M.; Ijumba, P.; Tsai, A.C.; Jackson, D. The use of mobile phones as a data collection tool: A report from a household survey in South Africa. *BMC Med. Inf. Decis. Mak.* **2009**, *9*, 1–8. [CrossRef] [PubMed]
3. McDonald, C. Big Data Opportunities for Telecommunications. Available online: https://mapr.com/blog/big-data-opportunities-telecommunications/ (accessed on 11 January 2019).
4. Telecom Regulatory Authority of India Highlights of Telecom Subscription Data as on 31 January 2019. Available online: https://main.trai.gov.in/sites/default/files/PR_No.22of2019.pdf (accessed on 21 February 2019).
5. Albreem, M.A.M. 5G wireless communication systems: Vision and challenges. In Proceedings of the 2015 International Conference on Computer, Communications, and Control Technology (I4CT), Kuching, SWK, Malaysia, 21–23 April 2015; pp. 493–497.

6. Weiss, G.M. Data Mining in Telecommunications. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2005; pp. 1189–1201.
7. Berson, A.; Smith, S.; Thearling, K. *Building Data Mining Applications for CRM*; McGraw-Hill Professional: New York, NY, USA, 1999.
8. Lu, H.; Lin, J.C.-C. Predicting customer behavior in the market-space: A study of Rayport and Sviokla's framework. *Inf. Manag.* **2002**, *40*, 1–10. [CrossRef]
9. Mendoza, L.E.; Marius, A.; Pérez, M.; Grimán, A.C. Critical success factors for a customer relationship management strategy. *Inf. Softw. Technol.* **2007**, *49*, 913–945. [CrossRef]
10. Hung, S.-Y.; Yen, D.C.; Wang, H.-Y. Applying data mining to telecom churn management. *Expert Syst. Appl.* **2006**, *31*, 515–524. [CrossRef]
11. Penders, J. Privacy in (mobile) Telecommunications Services. *Ethics Inf. Technol.* **2004**, *6*, 247–260. [CrossRef]
12. Agarwal, S.; Aulakh, G. TRAI Recommendations on Data Privacy Raises Eyebrows. Available online: https://economictimes.indiatimes.com/industry/telecom/telecom-policy/trai-recommendations-on-data-privacy-raises-eyebrows/articleshow/65033263.cms (accessed on 21 March 2019).
13. Hauer, B. Data and Information Leakage Prevention Within the Scope of Information Security. *IEEE Access* **2015**, *3*, 2554–2565. [CrossRef]
14. Blum, A.L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271. [CrossRef]
15. Lindell, Y.; Pinkas, B. Privacy Preserving Data Mining. In Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology, Santa Barbara, CA, USA, 20–24 August 2000; pp. 36–54.
16. Clifton, C.; Kantarcioğlu, M.; Doan, A.; Schadow, G.; Vaidya, J.; Elmagarmid, A.; Suciu, D. Privacy-preserving data integration and sharing. In Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD'04, Paris, France, 13 June 2004; pp. 19–26.
17. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, Georgia, 3–7 April 2006; p. 24.
18. Mendes, R.; Vilela, J.P. Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access* **2017**, *5*, 10562–10582. [CrossRef]
19. Karp, A.H. Using Logistic Regression to Predict Customer Retention. 1998. Available online: https://www.lexjansen.com/nesug/nesug98/solu/p095.pdf (accessed on 16 August 2019).
20. Mozer, M.C.; Wolniewicz, R.; Grimes, D.B.; Johnson, E.; Kaushansky, H. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Trans. Neural Netw.* **2000**, *11*, 690–696. [CrossRef]
21. Hur, Y.; Lim, S. *Customer Churning Prediction Using Support Vector Machines in Online Auto Insurance Service*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2005; pp. 928–933.
22. Larivière, B.; Van den Poel, D. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst. Appl.* **2005**, *29*, 472–484. [CrossRef]
23. Shao, J.; Li, X.; Liu, W. The Application of AdaBoost in Customer Churn Prediction. In Proceedings of the 2007 International Conference on Service Systems and Service Management, Chengdu, China, 9–11 June 2007; pp. 1–6.
24. Zhao, J.; Dang, X.-H. Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example. In Proceedings of the 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 12–17 October 2008; pp. 1–4.
25. Xie, Y.; Li, X.; Ngai, E.W.T.; Ying, W. Customer churn prediction using improved balanced random forests. *Expert Syst. Appl.* **2009**, *36*, 5445–5449. [CrossRef]
26. Lee, H.; Lee, Y.; Cho, H.; Im, K.; Kim, Y.S. Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) mode. *Decis. Support Syst.* **2011**, *52*, 207–216. [CrossRef]
27. Idris, A.; Khan, A.; Lee, Y.S. Genetic Programming and Adaboosting based churn prediction for Telecom. In Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea, 14–17 October 2012; pp. 1328–1332.

28. Idris, A.; Rizwan, M.; Khan, A. Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Comput. Electr. Eng.* **2012**, *38*, 1808–1819. [CrossRef]
29. Kirui, C.; Hong, L.; Cheruiyot, W.; Kirui, H. Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining. *Int. J. Comput. Sci. Issues* **2013**, *10*, 165–172.
30. Keramati, A.; Jafari-Marandi, R.; Aliannejadi, M.; Ahmadian, I.; Mozaffari, M.; Abbasi, U. Improved churn prediction in telecommunication industry using data mining techniques. *Appl. Soft Comput.* **2014**, *24*, 994–1012. [CrossRef]
31. Amin, A.; Shehzad, S.; Khan, C.; Ali, I.; Anwar, S. Churn Prediction in Telecommunication Industry Using Rough Set Approach. *New Trends Comput. Collect. Intell.* **2015**, *572*, 83–95.
32. Khodabandehlou, S.; Rahman, M.Z. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *J. Syst. Inf. Technol.* **2017**, *19*, 65–93. [CrossRef]
33. Erdem, K.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B.; Pentland, A.S. Behavioral attributes and financial churn prediction. *EPJ Data Sci.* **2018**, *7*, 1–18.
34. Amin, A.; Al-Obeidat, F.; Shah, B.; Adnan, A.; Loo, J.; Anwar, S. Customer churn prediction in telecommunication industry using data certainty. *J. Bus. Res.* **2019**, *94*, 290–301. [CrossRef]
35. Ahmad, A.K.; Jafar, A.; Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J. Big Data* **2019**, *6*, 1–24. [CrossRef]
36. Samarati, P.; Sweeney, L. Generalizing Data to Provide Anonymity when Disclosing Information. In Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Seattle, WA, USA, 1–4 June 1998; p. 188.
37. Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 571–588. [CrossRef]
38. Xu, J.; Wang, W.; Pie, J.; Wang, X.; Shi, B.; Fu, A.W.-C. Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 785–790.
39. Cormode, G.; Srivastava, D.; Yu, T.; Zhang, Q. Anonymizing bipartite graph data using safe groupings. *Proc. VLDB Endow.* **2008**, *1*, 833–844. [CrossRef]
40. Muntés-Mulero, V.; Nin, J. Privacy and anonymization for very large datasets. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 2117–2118.
41. Masoumzadeh, A.; Joshi, J. Preserving Structural Properties in Edge-Perturbing Anonymization Techniques for Social Networks. *IEEE Trans. Dependable Secur. Comput.* **2012**, *9*, 877–889. [CrossRef]
42. Emam, K.E.I.; Rodgers, S.; Malin, B. Anonymising and sharing individual patient data. *BMJ* **2015**, *350*, h1139. [CrossRef] [PubMed]
43. Goswami, P.; Madan, S. Privacy preserving data publishing and data anonymization approaches: A review. In Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 5–6 May 2017; pp. 139–142.
44. Bild, R.; Kuhn, K.A.; Prasser, F. SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees. *Proc. Priv. Enhancing Technol.* **2018**, *1*, 67–87. [CrossRef]
45. Liu, F.; Hua, K.A.; Cai, Y. Query l-diversity in Location-Based Services. In Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, Taipei, Taiwan, 18–20 May 2009; pp. 436–442.
46. Das, D.; Bhattacharyya, D.K. Decomposition+: Improving ℓ-Diversity for Multiple Sensitive Attributes. *Adv. Comput. Sci. Inf. Technol. Comput. Sci. Eng.* **2012**, *85*, 403–412.
47. Kern, M. Anonymity: A Formalization of Privacy-l-Diversity. *Netw. Archit. Serv.* **2013**, 49–56. [CrossRef]
48. Mehta, B.B.; Rao, U.P. Improved l-Diversity: Scalable Anonymization Approach for Privacy Preserving Big Data Publishing. *J. King Saud Univ. Comput. Inf. Sci.* **2019**, in press. [CrossRef]
49. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 16–20 April 2007; pp. 106–115.
50. Liang, H.; Yuan, H. On the Complexity of t-Closeness Anonymization and Related Problems. *Database Syst. Adv. Appl.* **2013**, *7825*, 331–345.

51. Domingo-Ferrer, J.; Soria-Comas, J. From t-Closeness to Differential Privacy and Vice Versa in Data Anonymization. *Knowl. Based Syst.* **2015**, *74*, 151–158. [CrossRef]
52. Soria-Comas, J.; Domingo-Ferrer, J.; Sánchez, D.; Martínez, S. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. In Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE), Helsinki, Finland, 16–20 May 2016; pp. 1464–1465.
53. Kumar, P.M.V. T-Closeness Integrated L-Diversity Slicing for Privacy Preserving Data Publishing. *J. Comput. Theor. Nanosci.* **2018**, *15*, 106–110. [CrossRef]
54. Evfimievski, A. Randomization in privacy preserving data mining. *ACM SIGKDD Explor. Newsl.* **2002**, *4*, 43–48. [CrossRef]
55. Aggarwal, C.C.; Yu, P.S. A Survey of Randomization Methods for Privacy-Preserving Data Mining. *Adv. Database Syst.* **2008**, *34*, 137–156.
56. Szűcs, G. Random Response Forest for Privacy-Preserving Classification. *J. Comput. Eng.* **2013**, *2013*, 397096. [CrossRef]
57. Batmaz, Z.; Polat, H. Randomization-based Privacy-preserving Frameworks for Collaborative Filtering. *Procedia Comput. Sci.* **2016**, *96*, 33–42. [CrossRef]
58. Kargupta, H.; Datta, S.; Wang, Q.; Sivakumar, K. Random-data perturbation techniques and privacy-preserving data mining. *Knowl. Inf. Syst.* **2005**, *7*, 387–414. [CrossRef]
59. Liu, L.; Kantarcioglu, M.; Thuraisingham, B. The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data. In Proceedings of the 6th IEEE International Conference on Data Mining, Hing Kong, China, 18–22 December 2006; pp. 507–512.
60. Shah, A.; Gulati, R. Evaluating applicability of perturbation techniques for privacy preserving data mining by descriptive statistics. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016; pp. 607–613.
61. Upadhyay, S.; Sharma, C.; Sharma, P.; Bharadwaj, P.; Seeja, K.R. Privacy preserving data mining with 3-D rotation transformation. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *30*, 524–530. [CrossRef]
62. Kotschy, W. The New General Data Protection Regulation—Is There Sufficient Pay-Off for Taking the Trouble to Anonymize or Pseudonymize data? Available online: https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf (accessed on 18 August 2019).
63. Stalla-Bourdillon, S.; Knight, A. Anonymous Data v. Personal Data—A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. *Wis. Int. Law J.* **2017**, *34*, 284–322.
64. Neumann, G.K.; Grace, P.; Burns, D.; Surridge, M. Pseudonymization risk analysis in distributed systems. *J. Internet Serv. Appl.* **2019**, *10*, 1–16. [CrossRef]
65. Telco Customer Churn Dataset. Available online: https://www.kaggle.com/blastchar/telco-customer-churn (accessed on 23 January 2019).
66. Tuv, E.; Borisov, A.; Runger, G.; Torkkola, K. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *J. Mach. Learn. Res.* **2009**, *10*, 1341–1366.
67. Mafarja, M.; Heidari, A.A.; Faris, H.; Mirjalili, S.; Aljarah, I. Dragonfly Algorithm: Theory, Literature Review, and Application in Feature Selection. *Nat. Inspired Optim.* **2019**, *811*, 47–67.
68. Mirjalili, S. Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Comput. Appl.* **2016**, *27*, 1053–1073. [CrossRef]
69. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
70. Biskup, J.; Flegel, U. Transaction-Based Pseudonyms in Audit Data for Privacy Respecting Intrusion Detection. In Proceedings of the Third International Workshop on Recent Advances in Intrusion Detection, London, UK, 2–4 October 2000; pp. 28–48.
71. Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption. Available online: https://www.xylem-technologies.com/2011/09/privacy-preserving-storage-and-access-of-medical-data-through-pseudonymization-and-encryption/ (accessed on 19 August 2019).
72. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18. [CrossRef]

73. Riza, L.S.; Nugroho, E.P. Metaheuristicopt: Metaheuristic for Optimization. Available online: https://cran.r-project.org/web/packages/metaheuristicOpt/metaheuristicOpt.pdf (accessed on 21 April 2019).
74. An R Package to Generate Synthetic Data with Realistic Empirical Probability Distributions. Available online: https://github.com/avirkki/synergetr (accessed on 23 May 2019).