Utilizing data sampling techniques on algorithmic fairness for customer churn prediction with data imbalance problems [version 2; peer review: 1 approved, 2 approved with reservations]

Maw Maw, Su-Cheng Haw, Chin-Kuan Ho

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

Abstract

Background: Customer churn prediction (CCP) refers to detecting which customers are likely to cancel the services provided by a service provider, for example, internet services. The class imbalance problem (CIP) in machine learning occurs when there is a huge difference in the samples of the positive class compared to the negative class. It is one of the major obstacles in CCP as it deteriorates performance in the classification process. Utilizing data sampling techniques (DSTs) helps to resolve the CIP to some extent.

Methods: In this paper, we review the effect of using DSTs on algorithmic fairness, i.e., to investigate whether the results pose any discrimination between male and female groups and compare the results before and after using DSTs. Three real-world datasets with unequal balancing rates were prepared and four ubiquitous DSTs were applied to them. Six popular classification techniques were utilized in the classification process. Both classifier’s performance and algorithmic fairness are evaluated with notable metrics.

Results: The results indicated that the Random Forest classifier outperforms other classifiers in all three datasets and, that using SMOTE and ADASYN techniques causes more discrimination in the female group. The rate of unintentional discrimination seems to be higher in the original data of extremely unbalanced datasets under the following classifiers: Logistics Regression, LightGBM, and XGBoost.

Conclusions: Algorithmic fairness has become a broadly studied area in recent years, yet there is very little systematic study on the effect of using DSTs on algorithmic fairness. This study presents important findings to further the use of algorithmic fairness in CCP research.

Keywords

Customer churn prediction, Data sampling techniques, Algorithmic fairness, Class imbalance problem
Corresponding author: Su-Cheng Haw (sucheng@mmu.edu.my)

Author roles: Maw M: Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation; Haw SC: Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; Ho CK: Project Administration, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is supported by the funding of TM Research & Development from Telekom Malaysia, Malaysia (Ref: MMUE/160013).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Maw M et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Maw M, Haw SC and Ho CK. Utilizing data sampling techniques on algorithmic fairness for customer churn prediction with data imbalance problems [version 2; peer review: 1 approved, 2 approved with reservations] F1000Research 2022, 10:988 https://doi.org/10.12688/f1000research.72929.2

First published: 30 Sep 2021, 10:988 https://doi.org/10.12688/f1000research.72929.1
In the revised version of the paper, we updated the introduction section with the applications of algorithmic fairness in two significant areas, recommender systems, and customer churn predictions. In addition, we also define the scope of the work. Three versions of data are prepared as extremely imbalanced (i.e., 5% non-churn and 95% churn rate), very imbalanced (i.e., 15% non-churn and 85% churn rate) and normal imbalanced (i.e., 30% non-churn and 70% churn rate). The classification results are demonstrated with the learning curves with the Random Forest classifier to provide a clear vision of the model's performance in these datasets and discussed briefly in the discussion section. We added a compact discussion on the major limitations of the study in the conclusion section. We believe this revised version of the paper would yield the readers with the insightful knowledge about the requirement to consider algorithmic fairness in machine-based decision tasks, especially in rapidly growing ecommerce businesses.

Any further responses from the reviewers can be found at the end of the article.

Introduction
Customer churn, the phenomenon in which customers are shifting to rival companies due to dissatisfaction with the existing services or for other inevitable reasons,1 is one of the common issues usually encountered in every customer-oriented sector, including telecommunication. Customer churn prediction (CCP) is a supervised binary classification procedure that detects the potential churners before they are churned. Since there are no standardized principles for collecting data for CCP tasks, data distribution between classes will be varied from one data set to another. Therefore, one class might have extremely underrepresented compared to another class. In CCP, the target class is those being churned or not. To be exact, churn is always a minority class when the non-churn class usually comes in large numbers. Therefore, churn is used to consider a rare object2 in service-based domains including telecom. Thus, telecom datasets always suffer from a class imbalance problem (CIP) and lead to a situation in which minority instances remain unlearned.

Advanced machine learning techniques can be applied to predict potential churners. Let us consider a dataset with 10,000 data instances with 10% of churn samples i.e., 1000 chunmers and 9,000 non-churners. Even if a carefully built model could predict 90% correctly on the minority class, it means 100 customers are misclassified to the wrong class. Suppose 60 chunmers are misclassified as non-chunmers, i.e., false negatives, the company will lose a huge amount of revenue since recruiting new customers is more expensive than keeping the existing ones.3 Thus, the ultimate goal in the telecom sector is to increase profit by decreasing customer churn. Hence, CIP is a block when trying to achieve the major goal of CCP, since it degrades classification accuracy. Algorithmic fairness has become a very active research topic since ProPublica observed that the algorithms could yield discriminative outcomes, which impacted a minority group in real life.4

Algorithmic fairness is monitored in line with the protected features or sensitive variables in the dataset. Sensitive data could be, but is not limited to gender, race, age group, or religion. Algorithmic fairness is achieved if the decisions generated by a model do not favor more or less any individual or a group.5 The lesser the bias in the training data, the bigger the chance of achieving algorithmic fairness. However, it is almost not possible to train a zero-bias model since the historical data could have contained bias for many reasons.6 The common reasons for bias in the training data involve the compounding of initial bias over time, using proxy variables, and unbalancing of sample size between minority and majority groups.7

In the CCP process, customers’ behavior is analyzed within specific time windows, for example within one month.8 Once the prediction is done, the outcomes are reused as training data for the next prediction. Therefore, there are high chance to have repeated bias in the historical data without even noticing. This fact suggests why the algorithmic fairness issue should be considered in building of CCP models. Since CIP is a major concern in CCP process, one solution for CIP is to apply data sampling techniques (DSTs) to the training data. Since the major function of DSTs is to increase or decrease the sample instances to balance between majority and minority classes, there are changes in the number of samples in the different groups in the dataset. Even though there are high chances of having bad impacts of using DSTs on algorithmic fairness, the research works on that specific topic are minimally found. This motivates us to start our research work on an evaluation of DSTs' impacts on algorithmic fairness in the CCP process.

In e-commerce businesses, the role of artificial intelligence has become very important, and it plays a major part to run the businesses more smoothly. On the other hand, along with the challenges of Covid-19, the trends of the customer-oriented businesses have rapidly transformed into the online platforms which need to heavily rely on the machine learning models to provide the precise personalized experiences to the customers. Thus, the works on preventing customer churns and enhancing recommendation systems are hot topics in the e-commerce research. Despite the research demand, there are very minimum works that are closely related to our study. Some of the prominent studies are organized in Table 1.
In this paper, we evaluate how the DSTs applied in the data preparation stage of the CCP process impact the algorithmic fairness by testing on three versions of real-world Telecom customer churn datasets. Therefore, the main goal of this study is to explore and identify the impact of using DSTs on training data on algorithmic fairness in the CCP process. To the best of our knowledge, there is very little research concerning algorithmic fairness in the CCP process. We believe the findings of this study would provide valuable insights into future CCP research.

In the following, we elaborate on the methodologies applied in the experimental procedure. We list all of the classification results and learning curve results, performance measures for fairness in the next section. The significant results are specifically discussed in the discussion section. After that, we discuss the challenges and opportunities gained from this study. We conclude the article by discussing the limitations and future works of the study.

Methods
Ethical Approval Number: EA1742021

Ethical Approval Body: Research Ethics Committee 2021, Multimedia University

In this study, the original data set is prepared to make three versions of unbalanced datasets, with rates of 5% (i.e., 95% of non-churn data and 5% churn data), 15%, and 30%. Each version is applied with four DSTs and compared the results with the unsampled original dataset to evaluate the classification performance and impacts on algorithmic fairness. The step-by-step methods to conduct the study are presented in Figure 1.

Datasets
A real-world telecom dataset was provided by one of Malaysia’s leading telecom companies (see Underlying data for details on access to this dataset). The original dataset contains 1,265,535 customer records, which were collected from January 2011 to December 2011. Since the original data set is huge in volume, we randomly selected 100,000 records and utilized them for this study as we prefer to start with a smaller sample of data due to the limited resource allocation. We included demographics, call information, network usage, billing information, and customer satisfaction data in our dataset since they are considered influential factors in the CCP process. A total of 22 features were extracted after careful aggregation, i.e., new features were created based on the original data and some unnecessary features were deleted from it, and features are listed in Table 2.

The final dataset was prepared with three different rates of unbalancing: 5%, 15%, and 30%. We created a Python script (see Extended data) which used the Pandas tool of Scikit-learn machine learning library to prepare three versions of datasets. We set up these specific rates because we wanted to experiment with extremely unbalanced cases up to intermediate levels.

| Research works | Description | Reference |
|----------------|-------------|-----------|
| Gui, Chun (2017) | In the article, the authors experimented with three kinds of feature selection techniques, Random Forest (RF), Relative Weight, Standardized regression coefficients, three types of DSTs, and RF classification algorithm on the real-world dataset with 45000 records to analyze the imbalanced dataset problem in CCP. | 3 |
| Mehrrotra et al. (2018) | This research work proposed a conceptual and computational framework and several policies for optimizing the relevance of the recommendation systems. | 9 |
| Valentim et al. (2019) | The authors evaluated the impact of different data preparation techniques mainly for sensitive data removal, encoding of categorical data, and data selection techniques by testing on two well-known public datasets. But their work was not focused on the DSTs’ impact on algorithmic and they tested only on very small public datasets. | 10 |
| Beutel et al. (2019) | The study introduced a new set of fairness evaluation metrics for the recommender systems by using pairwise regularization. | 11 |
| Patro et al. (2020) | The authors investigated the fair recommendation in the context of two-sided online platforms and proposed a novel mapping called FairRec which provides the effectiveness in incurring a marginal loss in the recommendation process. | 12 |
### Table 2. Features used in the real-world dataset.

| No. | Name of the features         | Description                                                                 |
|-----|------------------------------|-----------------------------------------------------------------------------|
| 1   | Customer ID                  | Customer ID                                                                 |
| 2   | Age                          | Age of customer                                                             |
| 3   | Is senior                    | Is the customer over 60 or not                                              |
| 4   | Gender                       | Gender of customer                                                          |
| 5   | Is local                     | Is the customer a Malaysian or an international?                            |
| 6   | Race                         | Is the customer Malay or Indian or Chinese or Other?                         |
| 7   | Technical-problem-count      | Total technical complaints and general complaints made by a customer        |
| 8   | Complain-count               | Total general complaints made by a customer                                 |
| 9   | Avrg download                | Average download rate                                                       |
| 10  | Avrg upload                  | Average upload rate                                                         |
| 11  | T-Location                   | The location where the customer registered for the service                  |
| 12  | HSBB area                    | Is the customer in the area where a high-speed connection is required or not |
| 13  | Speed                        | The broadband speed customer has registered for                            |
| 14  | Price start                  | The value of the package the customer has bought                            |
| 15  | Contract period              | The contract period of the customer                                         |
| 16  | Median- outstanding          | Average overdue fees                                                        |
| 17  | Avrg local amt               | Average amount spends on calling locally                                    |
| 18  | Avrg std amt                 | Average amount spends on subscriber trunk dialing                           |
| 19  | Avrg idd amt                 | Average amount spends on international calls                                 |
| 20  | Avrg voice usage             | Average amount spends on voice calls                                         |
| 21  | Avrg dialup amt              | Average amount spends on dialup service                                      |
| 22  | Churn                        | Whether the customer is churned or not                                       |
Data preprocessing

In the data preprocessing stage, we excluded any null values. Since we found only a few outliers in the selected dataset, we manually removed them without using any specific procedure. We applied four DSTs to the data: Random Over Sampler (ROS), Random Under Sampler (RUS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Oversampling Technique (ADASYN). The selection of DSTs was based on their popularity and to know the impact of each of them on the algorithmic fairness in the CCP process.

Classification of data

We applied six popular classifiers: Random Forest (RF), Decision Tree (DT), LightGBM (LGBM), Gradient Boosting (GB), Logistics Regression (LR), and XGBoost. We created our own Python script using Scikit-learn machine learning library to perform this step. After a careful exploratory data analysis, we dropped Customer ID, Avg local amt, Avg std amt, Avg idd amt, Avg dialup amt from the predictor variable list since they were weakly correlated to the target variable.

Evaluation of experiment

We performed two evaluations: performance measures and algorithmic fairness metrics.

Performance measures

In measuring the classifier's performance, we applied standard measures which are commonly used in most machine learning classification tasks, including precision, recall and accuracy. We applied F-1 and AUC-ROC scores since accuracy alone is not enough to evaluate the actual performance of the classifiers. However, the performances of the different classifiers are compared by using AUC-ROC score. We created our own script using Scikit-learn, a free machine learning software library for Python programming language. The performance of each classification was done as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

where

\[
TP = \text{true positive}
\]
\[
TN = \text{true negative}
\]
\[
FP = \text{false positive}
\]
\[
FN = \text{false negative}
\]

\[
\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}
\]

\[
\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}
\]

\[
\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{AUC - ROC} = \frac{\sum \text{Rank (+)} - |+| \times |(+)| + 1}{|+| + |-|}
\]

where

\[
\sum \text{Rank (+)} \text{ is the sum of all positive classified examples}
\]
\[
|+| \text{ is the number of positive examples in the dataset}
\]
\[
|-| \text{ is the number of negative examples in the dataset}
\]
Algorithmic fairness metrics

We emphasized the assessment of whether the classifier is discriminated against between women, a protected group, and men, a non-protected group. We applied two well-known fairness definitions in measuring algorithmic fairness and utilized the popular AI-fairness 360 tool to calculate algorithmic fairness.20

Statistical parity (SP): Also known as an equal acceptance rate. SP is achieved if women have an equal probability to be predicted in the positive, i.e., churn class, as the men.21

SP difference measures the difference of a specific outcome between the protected (female group) and non-protected (male group). The smaller the SP difference between the two groups, we can say that the model treats the unprotected group statistically similar to the protected group.

SP is calculated as follows:

\[
\Pr(Y = 1 \mid \text{Group} = \text{male}) = \Pr(Y = 1 \mid \text{Group} = \text{female}), \quad \text{where} \quad Y = \text{predicted decision}
\]

Disparate Impact (DI): Also known as indirect discrimination where no protected variables are directly applied, but biased outcomes are still produced relying on the variables correlated with protected variables.22 The standardized threshold for the calculation of DI is 0.8, which means the group whose DI values are under 0.8 are discriminated by the classifier.

The threshold value 80% is advised by the US Equal Employment Opportunity Commission.23 The model could be DI-free when the value is larger than 80% but it should be lower than 125%.24

DI is calculated as follows:

\[
DI = \frac{\Pr(Y = 1 \mid \text{Group} = \text{female})}{\Pr(Y = 1 \mid \text{Group} = \text{male})} \leq \tau = 0.8,
\]

where \( Y \) = predicted decision

Results

The preliminary classification results for the datasets with different data unbalanced rates using four DSTs are shown in Tables 3–5. Table 3 shows the specific results of classification performance gotten when testing on 5% of unbalanced rate concerning the chosen classifiers and four DSTs.

Table 4 shows the details results of classification performance obtained when testing on 15% of unbalanced dataset with respect to the chosen classifiers and four DSTs.

Table 5 shows the details results of classification performance obtained when testing on 30% of unbalanced dataset with respect to the chosen classifiers and four DSTs.

The classification results are shown in terms of learning curves in this section from Figures 2 to 4. In this article, we display the results of the classifications with RF since RF yields the best performance results among the other six classifiers. Figure 2 indicates the learning curves of the classification performance with 95%-5% (i.e, non-churn-churn) unbalanced rate of data compared with the original data and after four DSTs are applied.

Figure 3 reveals the learning curves of the classification performance with 85%-15% (i.e, non-churn-churn) unbalanced rate of data compared with the original data and after four DSTs are applied.

Figure 4 shows the learning curves of the classification performance with 70%-30% (i.e, non-churn-churn) unbalanced rate of data compared with the original data and after four DSTs are applied.

In our study, we have observed that a variable, is-senior remained unbalanced even after applying the DSTs. The algorithmic fairness scores for each group with different unbalanced rates are described in Tables 6–8. Table 6 shows the comparative results of SP difference and DI scores calculated on 5% unbalanced dataset and the original dataset.
Table 3. The classification results for the dataset with 5% unbalanced rate.

| Classifier | 5% imbalanced with ROS | | 5% imbalanced with RUS | | 5% imbalanced with SMOTE | | 5% imbalanced with ADASYN |
|-----------|-----------------------|---|-----------------------|---|-----------------------|---|-----------------------|
|           | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC |
| RF        | 0.99     | 0.99      | 0.99   | 0.99     | 0.99     | 0.84     | 0.84      | 0.84   | 0.84     | 0.93     | 0.83     | 0.83      | 0.83   | 0.83     | 0.93     | 0.83     | 0.83      | 0.83   | 0.83     | 0.93     |
| DT        | 0.98     | 0.98      | 0.98   | 0.98     | 0.97     | 0.80     | 0.80      | 0.80   | 0.80     | 0.79     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     |
| LGMB      | 0.89     | 0.89      | 0.89   | 0.89     | 0.96     | 0.84     | 0.84      | 0.84   | 0.84     | 0.93     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     |
| GB        | 0.85     | 0.85      | 0.85   | 0.85     | 0.99     | 0.85     | 0.85      | 0.85   | 0.85     | 0.93     | 0.85     | 0.85      | 0.85   | 0.85     | 0.85     | 0.85     | 0.85      | 0.85   | 0.85     | 0.85     |
| LG        | 0.78     | 0.81      | 0.78   | 0.78     | 0.87     | 0.80     | 0.83      | 0.80   | 0.81     | 0.88     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     |
| XGBoost   | 0.93     | 0.93      | 0.93   | 0.93     | 0.97     | 0.83     | 0.83      | 0.83   | 0.83     | 0.92     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     | 0.83     | 0.83      | 0.83   | 0.83     | 0.83     |

Table 4. The classification results for the dataset with 15% unbalanced rate.

| Classifier | 15% imbalanced with ROS | | 15% imbalanced with RUS | | 15% imbalanced with SMOTE | | 15% imbalanced with ADASYN |
|-----------|-----------------------|---|-----------------------|---|-----------------------|---|-----------------------|
|           | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC | Accuracy | Precision | Recall | F1-score | AUC-ROC |
| RF        | 0.97     | 0.97      | 0.97   | 0.97     | 0.99     | 0.76     | 0.75      | 0.76   | 0.76     | 0.83     | 0.75     | 0.75      | 0.75   | 0.75     | 0.83     | 0.75     | 0.75      | 0.75   | 0.75     | 0.83     |
| DT        | 0.93     | 0.93      | 0.93   | 0.93     | 0.93     | 0.68     | 0.68      | 0.68   | 0.68     | 0.68     | 0.68     | 0.68      | 0.68   | 0.68     | 0.68     | 0.68     | 0.68      | 0.68   | 0.68     | 0.68     |
| LGMB      | 0.78     | 0.78      | 0.78   | 0.78     | 0.87     | 0.76     | 0.77      | 0.76   | 0.76     | 0.84     | 0.77     | 0.77      | 0.77   | 0.77     | 0.84     | 0.77     | 0.77      | 0.77   | 0.77     | 0.84     |
| GB        | 0.76     | 0.77      | 0.76   | 0.76     | 0.84     | 0.76     | 0.76      | 0.76   | 0.76     | 0.84     | 0.76     | 0.76      | 0.76   | 0.76     | 0.84     | 0.76     | 0.76      | 0.76   | 0.76     | 0.84     |
| LG        | 0.64     | 0.64      | 0.64   | 0.64     | 0.69     | 0.64     | 0.64      | 0.64   | 0.64     | 0.70     | 0.64     | 0.64      | 0.64   | 0.64     | 0.70     | 0.64     | 0.64      | 0.64   | 0.64     | 0.70     |
| XGBoost   | 0.81     | 0.81      | 0.81   | 0.81     | 0.90     | 0.75     | 0.76      | 0.75   | 0.76     | 0.84     | 0.75     | 0.76      | 0.75   | 0.76     | 0.84     | 0.75     | 0.76      | 0.75   | 0.76     | 0.84     |
Table 5. The classification results for the dataset with 30% unbalanced rate.

| Classifier | 30% imbalanced with ROS | 30% imbalanced with RUS | 30% imbalanced with SMOTE | 30% imbalanced with ADASYN |
|------------|--------------------------|--------------------------|---------------------------|---------------------------|
|            | Accuracy  | Precision  | Recall  | F1-score | AUC-ROC  | Accuracy  | Precision  | Recall  | F1-score | AUC-ROC  | Accuracy  | Precision  | Recall  | F1-score | AUC-ROC  | Accuracy  | Precision  | Recall  | F1-score | AUC-ROC  |
| RF         | 0.89      | 0.89       | 0.89    | 0.89     | 0.95     | 0.74      | 0.74       | 0.74    | 0.74     | 0.82     | 0.86      | 0.86       | 0.86    | 0.92     | 0.92     | 0.86      | 0.86       | 0.86    | 0.92     | 0.92     |
| DT         | 0.83      | 0.83       | 0.83    | 0.83     | 0.83     | 0.66      | 0.66       | 0.66    | 0.66     | 0.66     | 0.79      | 0.79       | 0.79    | 0.79     | 0.79     | 0.79      | 0.79       | 0.79    | 0.79     | 0.79     |
| LGBM       | 0.76      | 0.76       | 0.76    | 0.76     | 0.84     | 0.75      | 0.75       | 0.75    | 0.75     | 0.83     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     |
| GB         | 0.74      | 0.75       | 0.74    | 0.74     | 0.82     | 0.75      | 0.75       | 0.75    | 0.75     | 0.83     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     |
| LG         | 0.66      | 0.67       | 0.66    | 0.66     | 0.68     | 0.63      | 0.64       | 0.63    | 0.63     | 0.68     | 0.79      | 0.79       | 0.79    | 0.79     | 0.79     | 0.79      | 0.79       | 0.79    | 0.79     | 0.79     |
| XGBoost    | 0.77      | 0.77       | 0.77    | 0.77     | 0.85     | 0.75      | 0.75       | 0.75    | 0.75     | 0.83     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     | 0.86      | 0.86       | 0.86    | 0.86     | 0.86     |
Table 7 displays the comparative results of SP difference and DI scores calculated on 15% unbalanced dataset and original dataset.

Table 8 describes the comparative results of SP difference and DI scores calculated on 30% unbalanced dataset and original dataset.

Discussion
Overview of experimental results
Recent works of algorithmic fairness research in machine learning applications is broadly organized into three main trends. Some studies emphasize enhancing or proposing better fairness notions and evaluation metrics in line with the domains concerned,21,25 some focus on the ways to mitigate the bias in the classification process (which can further be divided into three main groups: pre-, in-, and post-processing techniques),26–29 while the last trend proposes how to maintain the ethical AI standards and policies in practicing machine learning applications in different sectors.30,31

Despite some previous empirical studies on the impact of using preprocessing techniques on algorithmic fairness, the findings of previous works could not pinpoint the direct impact of using DSTs on algorithmic fairness. Lourenc and Antunes,10 which is the closest work to our research, distinguish the effect of data preparation on algorithmic fairness. However, their work has been tested with two small datasets and provides general results using random under- and over-DSTs. Importantly, their work fails to be tested on the widely-applied DSTs, SMOTE and ADASYN. In contrast, we apply real-world business data and show how different DSTs influence dissimilar levels of imbalance rate.

In the classification task, RF seems to be the best classifier since it yielded the best results over the other five models based on their AUC-ROC scores, while LR provided the worst scores for almost all metrics. It was observed that RUS worked better for the extremely unbalanced situation compared with 15% and 30% imbalanced rates. The best outcomes
were found via ROS, SMOTE, and ADASYN in all different unbalanced rates, thus, could be concluded that oversampling techniques seem to provide more promising prediction results than undersampling techniques. This might be because the undersampling technique modifies the data by decreasing the majority of instances, which makes the dataset lack useful information for learning. By observing the learning curves, the performance of the classifier is highly increased after applying DSTs in all versions of the data. But in the 5% unbalanced version, the performance results are better than two other versions of data. The classifier might ignore 5% of churns since the churn rate is too less. The data size used in the experiment is less, it would be better to use more training examples to see whether the models are well fit or not.

For all three unbalanced rates, the original dataset always gave less statistical parity differences (SPD) compared to sampled datasets created using four DSTs, while datasets with RUS and ROS yield a slightly larger SPD but the statistics showed there is no disparate impact. However, we can hypothetically consider there might still be a bias as both RUS and ROS have their limitations. With RUS, the important information could have been removed and the classifier could provide a biased result since there was less information to learn from. On the other hand, with ROS, the prediction performance could also be biased due to the overfitting problem. In this sense, it is suggestible to apply different fairness measures and to compare the fairness scores. For the DI scores, if there is DI less than 0.8, there is indirect discrimination towards the unprotected group. The mathematical equivalence of DI suggests equalizing the outcomes between protected and unprotected groups. However, in reality, the conditions in the context of interest drive us to allow DI to a specific group up to some percentage. For example, in telecom CCP, the number of female customers could be very less than the dataset, since most males usually apply for a network plan representing the whole household. Therefore, we assume considering DI with the 80% rule is reasonable.

In the 5% unbalanced original dataset, LGBM, LR and XG-Boost imposed with DI values of 0.79, 0.64, and 0.78 respectively. But there is no DI in the other two original datasets for 10% and 30%. This reveals that more discrimination could occur on a more unbalanced dataset. The analysis of all datasets with SMOTE and ADASYN provides alarming
information on the classifier’s discrimination against the unprotected group. The 30% unbalanced dataset yields the worst unfair results since this is the highest SPD between female and males groups with LR of 0.38 and 0.43, respectively. Overall, among all DSTs, ADASYN, and SMOTE tend to provide more unfair outcomes compared to other DSTs. In contradiction, they both provide a better classification performance in comparison to RUS and ROS. There is not a huge difference among the three different data unbalanced levels. However, in this study, we experimented with the gender attribute as a sensitive variable.

**Opportunities and challenges**

Due to the nature of the CCP process and the rarity issue, training datasets have high chances to have compounded bias and suffer from unbalanced problems not only for the target class but also in the other attributes including sensitive variables. We have noticed that one variable remained unbalanced even after applying the DSTs; in such a case, a careful selection of data attributes should be done to avoid selection bias.

As the quality of training data is important, we would suggest enhanced mechanisms of data repairing techniques to prevent bias in the training data. Furthermore, the algorithmic fairness problem mostly concerns societal discrimination. For example, in the scholarship selection process, if classifiers give more favors to males than females who have the same qualifications as males but are not selected, this will decrease their chances of a scholarship. In a profit-centered industry like telecom, one could think there will be no loss for the customers though any group is less or more favored. It is important to consider the impact of biased decisions for the sake of the company’s reputation, the importance of equal treatment to customers, and to practice ethical AI policies.

**Figure 4.** Before and after sampling results of learning curves with the data (70%-15% of non-churn and churn rate) with RF classifier.
Table 6. The algorithmic fairness measures on 5% unbalanced dataset.

| Algorithmic fairness metrics | 5% original data | 5% imbalanced with ROS | 5% imbalanced with RUS | 5% imbalanced with SMOTE | 5% imbalanced with ADASYN |
|-----------------------------|------------------|------------------------|------------------------|--------------------------|---------------------------|
| RF                          |                  |                        |                        |                          |                           |
| SP Difference               | -0.0056          | -0.0703                | -0.0524                | 0.1402                   | 0.1401                    |
| DI                          | 0.80             | 0.86                   | 0.89                   | 1.32                     | 1.32                      |
| LightGBM                    |                  |                        |                        |                          |                           |
| SP Difference               | -0.0067          | -0.0703                | -0.0524                | 0.1402                   | 0.1390                    |
| DI                          | 0.79             | 0.86                   | 0.89                   | 1.32                     | 1.32                      |
| GB                          |                  |                        |                        |                          |                           |
| SP Difference               | -0.0057          | -0.0854                | -0.5456                | 0.1349                   | 0.1319                    |
| DI                          | 0.80             | 0.84                   | 0.89                   | 1.31                     | 1.29                      |
| LR                          |                  |                        |                        |                          |                           |
| SP Difference               | -0.0024          | -0.0789                | -0.0227                | 0.1765                   | 0.1591                    |
| DI                          | 0.64             | 0.94                   | 0.96                   | 1.34                     | 1.28                      |
| XGBoost                     |                  |                        |                        |                          |                           |
| SP Difference               | -0.0069          | -0.0842                | -0.5446                | 0.1387                   | 0.1383                    |
| DI                          | 0.78             | 0.85                   | 0.89                   | 1.32                     | 1.32                      |
| DT                          |                  |                        |                        |                          |                           |
| SP Difference               | -0.0073          | -0.0698                | -0.0397                | 0.1331                   | 0.1317                    |
| DI                          | 0.86             | 0.87                   | 0.91                   | 1.30                     | 1.29                      |

Table 7. The algorithmic fairness measures on 15% unbalanced dataset.

| Algorithmic fairness metrics | 15% original data | 15% imbalanced with ROS | 15% imbalanced with RUS |
|-----------------------------|-------------------|-------------------------|-------------------------|
| RF                          |                   |                         |                         |
| SP Difference               | -0.009644         | -0.058223               | -0.049729               |
| DI                          | 0.871212          | 0.892587                | 0.891628                |
| LightGBM                    |                   |                         |                         |
| SP Difference               | -0.011044         | -0.070341               | -0.059666               |
| DI                          | 0.856964          | 0.853293                | 0.872208                |
| Gradient Boosting           |                   |                         |                         |
| SP Difference               | -0.006168         | -0.057517               | -0.055826               |
| DI                          | 0.913621          | 0.874225                | 0.877351                |
| Logistic Regression         |                   |                         |                         |
| SP Difference               | -0.007872         | -0.1004                 | 0.01865                 |
| DI                          | 0.711771          | 0.79581                 | 1.03868                 |
| XGBoost                     |                   |                         |                         |
| SP Difference               | -0.011008         | -0.069346               | -0.059144               |
| DI                          | 0.864076          | 0.858386                | 0.874791                |
| Decision Tree               |                   |                         |                         |
| SP Difference               | -0.016556         | -0.05406                | -0.016812               |
| DI                          | 0.900914          | 0.907414                | 0.967054                |
Conclusions
In this paper, we experimented on three versions of unbalanced real-world telecom datasets to assess the impact of using four types of DSTs on the algorithmic fairness in the CCP process and compared the results with the unsampled original dataset. Classification performance and algorithmic fairness were evaluated with well-known metrics. The outcomes imply that RF provides the best classification results. Using SMOTE and ADASYN yields larger SPD between male and female groups as well as a disparate impact on the female over the male group. Previous work emphasizes the use of this method in choosing a scholarship candidate, releasing prisoners on parole, and choosing a credible candidate. Since machine learning applications would be applied to almost every sector so on shortly, the practice of using fairer or unbiased systems is essential. Our study highlights the importance of paying attention to algorithmic fairness in the machine-driven decision-making process of the profit-centered and customer-oriented sectors on which very little research work has been done. Particularly, our finding highlights the fact that a careful choice of DSTs must be done to achieve unbiased prediction results. Despite the experiment being done on the real-world telecom dataset with 100,000 records, the data size is still considered to be very less and the experimental results on a much larger amount of data could be varied. Since the data size is small, the results on the learning curves are not remarkable. Another limitation is we use only one sensitive attribute, gender, in our study. In future work, we would like to test the same procedure on a much larger dataset and would like to measure more algorithmic fairness metrics to investigate the best suitable algorithmic measures for the CCP task. Moreover, we would like to test with more sensitive variables rather than just gender.

Data availability
Underlying data
The real-world telecom dataset was obtained from the Business Intelligence and Analytics department of Telekom Malaysia Bhd. The authors were required to go through a strict approval process following established data governance framework. Interested readers/reviewers may contact the Business Intelligence and Analytics department to request the data (technicalsupport@tm.com.my). The decision as to whether or not to grant access to the data is at the discretion of Telekom Malaysia Bhd.

As most telco companies own similar customer data, other customer churn datasets that are representative of the data being used in this research can be found as follows:
1. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn.

2. https://datasetsearch.research.google.com/search?query=Telco%20Customer%20Churn%20dataset%20site%3Akaggle.com&docid=L2cvMTFsbDF0dzJ5NA%3D%3D.

Extended data
Analysis code available from: https://github.com/mawmaw/fairness_churn.

Archived analysis code as at time of publication: https://doi.org/10.5281/zenodo.5516218.

License: MIT License.

References

1. Eria K, Markarian BP: Systematic Review of Customer Churn Prediction in the Telecom Sector. 2018; vol. 2(0), 1.
2. Amin S, Arif, et al.: Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. IEEE Access. 2016; 4(MI): 7940–7957. Publisher Full Text
3. Gui C: Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. 2017; vol. 4(2), pp. 93–99.
4. Dietrich W, Mendoza C, Brennan T: COMPAQS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. 2016.
5. Nikouli E, et al.: Bias in Data-driven AI Systems – An Introductory Survey. 2020; pp. 1–19.
6. Kamiran F, Ziogetaitė I: Explainable and non-explainable discrimination in classification. Stud. Appl. Philos. Epistemol. Rational. Ethics. 2013; 3(January 2012): 155–170. Publisher Full Text
7. Barocas S, Selbst AD: Big Data’s Disparate Impact. 2016; vol. 671: pp. 671–732.
8. Ballings M, Van Den Poel D, Verhagen E: Improving customer churn prediction by data augmentation using pictorial stimulus-choice data. Adv. Intell. Syst. Comput. 2012; 171(1): 217–226. Publisher Full Text
9. Mehrora R, McInerney J, Bouchard H, et al.: Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. Int. Conf. Inf. Knowl. Manag. Proc. 2018: 2243–2252.
10. Lourenc N, Antunes N: The Impact of Data Preparation on the Fairness of Software Systems.
11. Beutel A, et al.: Fairness in recommendation ranking through pairwise comparisons. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2019: 2212–2220.
12. Patro GK, Biswas A, Ganguly N, et al.: FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. Web Conf. 2020; Proc. World Wide Web Conf. WWW 2020: 2020; 2: 1194–1204.
13. Culemelli L, Nunez-Del-Prado M, Zarate-Gamarra L: Measuring churner influence on pre-paid subscribers using fuzzy logic. Proc. 2016 42nd Lat. Am. Comput. Conf. CLEI 2016.
14. Ahmed U, Khan A, Khan SH, et al.: Transfer Learning and Meta Classification Based Deep Churn Prediction System for Telecom Industry. 2019; pp. 1–9.
15. Mohammed R, Rawashdeh J, Abdullah M: Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020. 2020: 243–248.
16. Chawla NV, Bowyer KW, Hall LO: SMOTE: Synthetic Minority Over-sampling Technique. 2002; 16: 321–357.
17. Gosain A, Sardana S: Handling class imbalance problem using oversampling techniques: A review. 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017. 2017; January: 79–85.
18. Kotiantis SB: Supervised Machine Learning: A Review of Classification Techniques. Informatica. 2007; 31(2007): 249–268.
19. Kotu V, Deshpande B: Model Evaluation. Predict. Anal. Data Min. 2015; 257–273. Publisher Full Text
20. Bellamy RKE, et al.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. 2018.
21. Dwork C, Hardt M, Pitassi T, et al.: Fairness through awareness. ITCS 2012 - Innov. Theor. Comput. Sci. Conf. 2012; 214–226.
22. Calmon FP, Wei D, Ramamurthy KN, et al.: Optimized Data Pre-processing for Discrimination Prevention. 2017; 1–18.
23. Feldman M, Friedler SA, Moeller J, et al.: Certifying and removing disparate impact. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2015; 2015-August: 253–268.
24. Zafar MB, Valera I, Goel S, et al.: Fairness Constraints: Mechanisms for Fair Classification. 2017; 54.
25. Hardt M, Price E, Srebro N: Equality of opportunity in supervised learning. Adv. Neural Inf. Proces. Syst. 2016; 3323–3331.
26. Zemel R, Wu Y, Swersky K, et al.: Learning fair representations. 30th Int. Conf. Mach. Learn. ICML 2013; 2013; 28(PART 2): 1362–1370.
27. Calmon FP, Wei D, Vázquez MA, et al.: Optimized pre-processing for discrimination prevention. Adv. Neural Inf. Procs. Syst. 2017; 2017-Decem(Nips): 3993–4002.
28. Zhang BH, Maboudian H, Mitchell M: On fairness and calibration. Adv. Neural Inf. Procs. Syst. 2017; 3331.
29. Pleiss G, Raghavan M, Wu F, et al.: On fairness and calibration. Adv. Neural Inf. Procs. Syst. 2017; 2017-Decem(Nips): 5681–5690.
30. Gursoy ME, Tamersoy A, Truex S, et al.: Secure and Utility-Aware Data Collection with Condensed Local Differential Privacy. IEEE Trans. Dependable Secur. Comput. 2019; X: 1–1.
31. Hube C, Athanas H, Gadiraju U: Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. Conf. Hum. Factors Comput. Syst. - Proc. 2019.
32. maxmaw: mawmaw/fairness_churn: Initial (v1.0). Zenodo. 2021. Publisher Full Text
DSTs (one under-sampling and three over-samplings) were applied to the three datasets. However, it would be clear for the readers if the authors could specify how many percent of the major class was under-sampled and how many percent of the minor class was over-sampled.

It is observed that a variable, "is-senior", remained unbalanced even after applying the DSTs. Please explain whether the DSTs (under-sampling or over-sampling?) were applied to the whole dataset dimension or some specific variables only.

It is stated that the under-sampling technique modifies the data by decreasing the majority of instances, which makes the dataset lack useful information for learning. But this should benefit the classification algorithm as the impact from the major class reduces. The authors can probably explain from the perspective of the rarity of the churn class.

For the tables, please state the reason certain numbers are bolded. Suggest to put the reason in the caption.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes
Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data Mining and Computer Networks

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1
Reviewer Report 24 February 2022
https://doi.org/10.5256/f1000research.76542.r120732

© 2022 P P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Prabu P
Department of Computer Science, CHRIST University, Bengaluru, Karnataka, 560029, India

This work mainly focuses on data imbalance problems in customer churn prediction. The author has to incorporate the following suggestions in his/her article in order to improve the quality of the work:

1. The author needs to clearly define performance metrics for comparing random forest classifier with other classifiers.

2. More experimental studies required for unequal balancing datasets and author expected to try the experiment with large datasets.

3. Author need to identify some proper technique to manage outliers in the datasets instead of removing manually.

4. Results and Algorithms fairness need to be explain properly.

5. Introduction section are too short and the section is not continuous. Correct it. The introduction section is not clearly explaining the basics and also lack of some details of about "issues in existing work, scope of work, goal of work and organization of work" can be included.

6. No significant limitations are discussed. A number of limitations and learning points are
also be considered after the conclusion.

7. Author could take more recent research papers for the literature review and most of the literature review papers are not suitable to the proposed work. Author is expected to include few existing recommender system suggested for e-commerce platform. So I would suggest the author should identify few more recent papers based on his proposed research work and also include the objective and limitations of each work in table format.

**Is the work clearly and accurately presented and does it cite the current literature?**
Yes

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Machine Learning, Cloud Computing ad Mathematical Modeling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 Jun 2022

Haw Su Cheng , Multimedia University, Cyberjaya, Malaysia

1. AUC-ROC score was applied to compare the performance of the respective classifiers and this information is added for clarification in the discussion section/ performance measure section

2. Due to the dataset availability on propriety company, the focus on this paper is on 100,000 records only. Testing with large dataset is put as future work. In future work, we would like to test the same procedure on a larger dataset and would like to measure more algorithmic fairness metrics to investigate the best suitable algorithmic measures for the CCP task.
3. Since there were not a lot of outliers in the original dataset, they were manually removed and we added this fact in the data pre-processing section.

4. In our study, we do not go further process of mitigating algorithmic unfairness yet. If so, we will need to apply and discuss. But here, we only evaluate and compare the bias in the training dataset before and after applying the data sampling techniques.

5. Introduction section is updated with related works, scope of work and organization work. Since my background is the combination of three different topics, algorithmic fairness, customer churn prediction and data sampling techniques, Due to the page limitation, could not put a lot of materials in the introduction section but most important and relevant background study was provided in the introduction section in a compact way. Goal of work is provided in the last paragraph of the introduction section, “The main goal of this study is to explore and identify the impact of using DSTs on training data on algorithmic fairness in the CCP process.”

6. We updated with the limitations in the conclusion section as per suggestion.

7. Although there is very limited work which are related to my work, we organize some of the articles of fairness in recommender system in the table format as per your suggestion

**Competing Interests:** No competing interests were disclosed.
Is the work clearly and accurately presented and does it cite the current literature?  
Yes

Is the study design appropriate and is the work technically sound?  
Yes

Are sufficient details of methods and analysis provided to allow replication by others?  
Yes

If applicable, is the statistical analysis and its interpretation appropriate?  
Yes

Are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions drawn adequately supported by the results?  
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

Author Response 06 Jun 2022

**Haw Su Cheng**, Multimedia University, Cyberjaya, Malaysia

Learning curves for Random Forest before and after applying DSTs (for three versions of datasets) are provided. We discussed briefly in the discussion section as well.

**Competing Interests:** No competing interests were disclosed.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com