國立陽明交通大學

資訊管理研究所

碩士論文

Institute of Information Managemant

National Yang Ming Chiao Tung University

Master Thesis

以欠採樣及過採樣降低機器學習中的偏誤

Under-Sampling and Over-Sampling for Debias in Machine Learning

研究生: 康佑誠 (Kang, Yu-Cheng)

指導教授：陳柏安 (Chen, Po-An)

中華民國一一一年八月

August 2022

以欠採樣及過採樣降低機器學習中的偏誤

Under-Sampling and Over-Sampling for Debias in Machine Learning

研 究 生：康佑誠
指導教授：陳柏安 博士

Student: Yu-Cheng Kang
Advisor: Dr. Po-An Chen

國立陽明交通大學

資訊管理研究所

碩 士 論 文

A Thesis
Submitted to Institute of Information Managemant
College of Management
National Yang Ming Chiao Tung University
in partial Fulfilment of the Requirements
for the Degree of
Masterof Science
in
Information Management

August 2022
Taiwan, Republic of China

中 華 民 國 一 一 一 年 八 月

以欠採樣及過採樣降低機器學習中的偏誤

學生：康佑誠　　指導教授：陳柏安 博士

國立陽明交通大學 資訊管理研究所

摘　　要

機器學習中存在著許多的歧視，也就是這些不公平的決策，可能偏頗某些優勢族群。造成這些歧視的原因可能有很多種，其中又包括資料標籤中本身的歧視 (偏見歧視): 人類決策主觀的歧視對標籤結果的影響、以及取樣的偏誤: 因爲不一致的採樣策略導致某些族群的有利結果代表性不足，也視爲類別不平衡的問題。

在這篇論文，我們先後探討現今主要對於機器學習中歧視的減緩方式，包含前處裡、中處理以及後處理；接著我們在先前提出資料增強的研究，也就是對資料進行過採樣生成，以減緩訓練集資料中類別不平衡的技術基礎下，先使用基本的隨機過採以及隨機欠採技術來證實對於訓練集資料進行採樣提升公平性的效用。在基於這個基礎之下，接著提出使用現有的結合過採樣以及欠採樣兩種組合採樣技術，分別是 Smote+ENN 以及 Smote+TomekLinks，用於平衡資料類別不平衡的情況且移除資料中被視爲噪點及邊界值的樣本；另外針對掌握控制群體分布比例的能力，我們改進了 TomekLinks 欠採樣演算法。最後在幾個公開資料集中，實驗結果顯示結合過採樣及欠採樣的技術相較於先前的研究，在公平性上皆略有所提升。

關鍵字: 機器學習偏見、組合採樣增強、AI 公平性、集成學習、SHAP

Under-Sampling and Over-Sampling for Debias in Machine Learning

Student: Yu-Cheng Kang Advisor: Dr. Po-An Chen

Institute of Information Managemant
National Yang Ming Chiao Tung University

Abstract

There are various prejudice、discrimination in machine learning, that is, these unfair decisions may be biased against certain unprivileged groups. There may be many reasons for these discrimination, including the discrimination in the data label itself (prejudice bias): the influence of subjective discrimination of human decision-making on the labeling results, and the sampling bias: it is due to inconsistent data collection strategy resulting in underrepresentation of a certain group in the dataset, which is considered a class imbalance problem.

In this thesis, we discussed the main ways to mitigate discrimination in machine learning nowadays, including pre-processing, in-processing, and post-processing. Based on the data augmentation techniques we have previously surveyed, that is, oversampling data to reduce class imbalance in the training data sets, we first use random oversampling and random undersampling techniques to verify the effectiveness of sampling technique on improving fairness in the data. Based on this foundation, we further proposed using the existing combination sampling technique, which combines oversampling and undersampling techniques, such as SMOTE+ENN and SMOTE+TomekLinks, to mitigate data class imbalance. In addition, in order to gain control of group distribution ratio, we modified the TomekLinks undersampling algorithm. In several public datasets, the results of the experiment show that, compared to previous studies, the combination of oversampling and undersampling techniques can slightly improve fairness while keeping a well accuracy.

Keywords: Machine Learning Bias, Combination Sampling Augmentation, AI Fairness, Ensemble Learning, SHAP

Acknowledgement

首先這篇論文能夠順利的完成，首先要感謝的是我的指導教授陳柏安老師。在我心中柏安老師對學術研究有著極大的熱忱，對待學生也非常有耐心；從碩一開始就組了一個讀書會一起和我們研讀論文新知，直到碩二開始尋找題目、相關論文，一路上都是老師一步步帶領我們探索研究的方向以及學習新知；在最後的實驗過程、論文撰寫上老師也提供了許多建議與協助。也要謝謝陳豐奇老師在和我們共同 meeting 時也給了許多不同的見解，以及口委郭柏志老師、李永銘老師給的建議，讓這篇論文能更加完善。

我也要謝謝 EC Lab 朝夕相處的夥伴們: 敦凱、黃靖、鈺蓁、哲維、耀云、可爲、林晢，還有 OTA、NET Lab 幾個比較熟的同學，一起在實驗室裡度過無數個熬夜趕期中考、期末專案，以及在找實習時大家互相的幫忙、建議，還有固定禮拜四的打排球時間，是我們在繁忙碩士生活當中共聚的美好時光；謝謝你們爲我的碩士生涯增添色彩，也留下許多難忘的回憶。

最後我要感謝我的家人們，默默地在精神、經濟上的支持我，讓我能無後顧之憂地專心攻讀碩士；也要謝謝一路上曾經幫助過我的同學、學長姐、朋友們。

康佑誠　謹誌

國立陽明交通大學資訊管理所

中華民國一一一年八月

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial intelligence, or more precisely, machine learning models, is used by many private enterprises, government agencies, and non-profit organizations to assist human decision-making processes, such as credit card scoring, movie recommendation, and crime recidivism assessment nowadays. [1]

However, in the process from data collecting, model training to decision making, various biases creep in. Even in the field of medical imaging, there are also bias and discrimination issues. In [2], the authors mentioned that the classification performance of the AI system will vary between different demographic groups. If such a model is used in the medical field, it may lead to black or female patients being more likely to be wrongly identified as healthy than those who are white or male. This is a risky and costly problem in the medical field.

The source of bias can be roughly divided into two types: Prejudice bias and Sample Bias (incomplete or unrepresentative training data) [1] [3]. Prejudice bias could be caused by cultural influences or stereotypes. Social class, status, race, nationality, gender, or even appearances can creep into the ML model, which can unjustly skew the results. However, sampling bias exists when a certain group is incomplete or unrepresentative in the training data. This could be due to the inconsistent or flawed data collection process.

We can divide fairness into two types according to the definition: Individual fairness and Group fairness. Individual fairness requires that each individual in the same group should receive the same treatment. [4] Common individual fairness metrics include consistency score and theil index. Group fairness is also known as statistical parity. It requires the protected group (e.g. unprivileged group) to be treated the same as the opposite group (e.g. privileged group) or the overall group.

The measurement indicators commonly used to evaluate group fairness include Statistical Parity Difference (SPD), Disparate Impact (DI), Average Odds Difference (AOD), Average odds error(AOE) and Equality Opportunity Difference (EOD), etc.

With these indicators for measuring bias, a lot of methods for mitigating bias have also been proposed one after another, which can be roughly divided into three types: pre-processing, in-processing, and post-processing.

In this thesis, we will focus our work mainly on group fairness, and we plan to refer to the data augmentation technique proposed by Sharma et al., 2020 [3], and lossless de-biasing proposed by Zhou et al., 2021 [5] on four well-known opensource datasets such as Adult Census, COMPAS, German Credit, and Bank datasets to reduce bias. Both of these two are mitigating bias methods that belong to pre-processing. We first experiment with the basic sampling methods: random oversampling and random undersampling to verify the validity and robustness of the formula justification in [5] to obtain a fairer result.

And then, based on the above two pre-processing methods, we introduce using two sampling combination techniques proposed by Gustavo et al., 2017 [6]: SMOTE + ENN and SMOTE + Tomek links, on the open source datasets mentioned above. In addition, since Tomek links as a cleaning data sample cannot maintain the originally predetermined ratio, we have made some modification to the Tomek links algorithm so that the ratio after sampling can meet our requirements.

Experiments show a promising result that these combination sampling techniques which combine oversampling and undersampling techniques can reduce the data imbalance in the training datasets, which is considered the reason for causing bias, and consequently mitigate the bias and obtain a fairer outcome.

Our main contributions in this thesis are as follows:

1. We propose to use a combination of oversampling and undersampling as a preprocessing technique to remove bias in machine learning.

2. We modify the Tomek links algorithm and make the modified version maintain our desired population ratio for the dataset.

1.1 Related Work

In order to mitigate bias in the machine learning model to meet the fairness objectives, the methods can be roughly divided into three types:

1.1.1 Pre-processing

In pre-processing, training data will be processed before being fed into the machine learning model to reduce discrimination or bias in the data. Common pre-processing techniques include optimized pre-processing which is proposed by Calmon et al.[7], reweighting proposed by Kamiran and Calders[8], the disparate impact remover designed by Feldman et al.[9], and data augmentation technique proposed by Zhou et al.[5] and Sharma et al.[3] aims to synthesize data to create new datasets to meet fairness objectives and mitigate bias.

Optimized preprocessing is a technique that learns and applies probabilistic transformation adjustments to features and labels of data through convex optimization while satisfying group fairness, individual distortion, and data utility constraints[7]. But in the optimized pre-processing technique, test data need to be transformed before being fed into the classification model, which is a little bit different from the data augmentation technique we implement in this thesis.

In Sharma et al.[3], they propose a new data augmentation technique to mitigate the bias in the data. What they do is: for each sample in the dataset, create a new sample with the same attributes, but with the opposite value for the protected attribute. These synthesized samples are first sorted in descending order according to the proximity to the distribution of their original data, and then added to the original data in order successively.

1.1.2 In-processing

In-processing refers to adjusting the way a machine learning model is trained to reduce bias. For some reasons such as proprietary or licensed data being hard to access and restricted to modify, pre-processing is hence infeasible. When this happened, we can consider using in-processing as a way to mitigate bias. Common in-processing techniques include the adver-

sarial debiasing method proposed by Zhang et al.[10] and the Prejudice Remover designed by Kamishima et al.[11].

1.1.3 Post-processing

Post-processing techniques modify output labels to meet different fairness objectives.[5] When we do not get the chance or ability to access the training data or to modify the algorithm the model used, post-processing may be a good choice for us. Reject option proposed by Kamiran et al.[12] is one of the most well-known methods of post-processing.

Chapter 2

Preliminaries

In **preliminaries**, we will first discuss the definitions of fairness measurements that we will use in this thesis, including group fairness and individual fairness. Next, we will discuss the oversampling and undersampling methods used in this thesis and their combinations. Finally, we will introduce the lossless de-biasing using oversampling proposed by Zhou's[5], which is also one of the methods we refer to in this thesis.

2.1 Fairness Measurements

Many previous studies have proposed several indicators to measure the fairness of the machine learning model, that is, to see whether the machine learning model is discriminatory and therefore provide a way that can be used to measure the effectiveness of mitigating bias.

Before discussing metrics for measuring machine learning fairness, let us distinguish two types of fairness: Group fairness and Individual fairness. Next, we will introduce the metrics we use in this thesis under the two types of fairness definition.

2.1.1 Individual Fairness Measurement

At the level of individual fairness, machine learning is often required to treat similar individuals equally. From a mathematical point of view, we can usually represent an individual through a set of parameters in a multidimensional space, and in classification tasks, adjacent sample points under the same sample parameter space will be more likely to obtain the same prediction result from the machine learning algorithm.

- **Consistency Score** : Individual fairness metric from [13] that measures how similar the labels are for similar examples. A higher score means higher chances that the same indi-

vidual will receive the same treatment.

$$ConsistencyScore = 1 - \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - \frac{1}{n\_neighbors}\sum_{j\in N_n\_neighbor(xi)}\hat{y}_j| \quad (2.1)$$

The Consistency Score uses the KNN algorithm to find the neighbors of each sample, and then compares its label with their neighbors'.

2.1.2 Group Fairness Measurement

On the other hand, at the level of group fairness, starting from a statistical point of view, we will divide the overall population into subgroups and observe the gap in the positive rate between the subgroups or take into account the classification performance factor: observe their gap in true positive rate, false positive rate, or both.

In group fairness, these gaps are generally required to be as small as possible. And group fairness is also a measurement that most of the current research on machine learning bias focuses on. The following are the group fairness metrics that we will use in this thesis, and these metrics also correspond to three mathematical criteria: Independency, Separation, and Sufficiency.

- **Statistical Parity Difference (SPD)** : SPD measures the gap between favorable outcomes for privileged and unprivileged groups.

$$\begin{aligned} SPD &= PR_U - PR_P \\ &= Pr(Y' = 1|A = 0) - Pr(Y' = 1|A = 1) \end{aligned} \quad (2.2)$$

- **Disparate Impact(DI)** : DI is the ratio of favorable outcomes rate of the unprivileged group to the privileged group.

$$\begin{aligned} DI &= PR_U/PR_P \\ &= Pr(Y = 1|A = 0)/Pr(Y = 1|A = 1) \end{aligned} \quad (2.3)$$

What SPD and DI pursues can correspond to Independency:

$$Y' \perp A$$

To satisfy Independency criteria, the classifier's prediction result should be independent of the sensitive attribute(protected attribute such as gender, age or race) the model may be biased. This can also be understood as that the classifier prediction results and the sensitive attribute have no mutual information.

- **Equal Opportunity Difference (EOD)** : EOD measures the gap of true positive rates between the privileged and the unprivileged groups.

$$\begin{aligned} EOD &= TPR_U - TPR_P \\ &= Pr(y' = 1|Y = 1, A = 0) - Pr(Y' = 1|Y = 1, A = 1) \end{aligned} \tag{2.4}$$

And it can correspond to the definition of Separation:

$$Y' \perp A|Y$$

To satisfy Separation criteria, the classifier's prediction result should be independent of the sensitive attribute given the condition on the true label.

- **Average Odds Difference (AOD)** : AOD is used to measure the mean difference in the true positive rate and the false positive rate between the privileged group and the unprivileged group.

$$\begin{aligned} AOD &= 0.5 * ((TPR_U - TPR_P) + (FPR_U - FPR_P)) \\ &= 0.5 * (Pr(Y' = 1|Y = 1, A = 0) \\ &\quad - Pr(Y' = 1|Y = 1, A = 1) \\ &\quad + Pr(Y' = 1|Y = 0, A = 0) \\ &\quad - Pr(Y' = 1|Y = 0, A = 1)) \end{aligned} \tag{2.5}$$

AOD can correspond to the definition Sufficiency:

$$Y \perp A|Y'$$

To satisfy Sufficiency criteria, the true label should be independent of the sensitive attribute given the condition on the classifier's prediction result.

$$\begin{aligned} PR &= (TP+FP)/(TP+FP+TN+FN) \\ TPR &= TP/(TP+FN) \\ FPR &= FP/(FP+TN) \end{aligned} \tag{2.6}$$

Where $A = 1$ stands for the privileged group, $A = 0$ stands for the unprivileged group, and $Y = 1$ represents the favorable outcome, $Y = 1$ represents the unfavorable outcome, and Y' is the prediction result.

How we choose the appropriate metric depends on our own preference for the outcome of the task or according to our moral and legal requirements. For example, in terms of college admissions or corporate recruitment, we usually hope or require that the admission rates between different groups be closer to achieve demographic parity. At this time, SPD or DI can be used. In terms of bank loans, banks may focus more on the gap in successful loans between different subgroups to meet the equality of opportunity. Then users like banks can use EOD as a measurement. On the other hand, in the case of medical diagnostic testing, because they want the test to perform equally well on individuals in each subgroups[14], then AOD will be their top choice.

2.2 Under-Sampling, Over-Sampling and their Combination

Common methods for dealing with class imbalance include oversampling and undersampling. In Gustavo et al., 2017[6], the influence of class imbalance on classifier performance was discussed. The authors found that in addition to class imbalance, the problem of class over-

lap in data can also cause performance impact, and based on the above findings, the authors proposed a combination of oversampling and undersampling techniques (SMOTE+ENN and SMOTE+Tomek inks) not only to deal with the problem of class imbalance but also to remove the noisy example between the decision boundary to help find data that better define the decision boundary.

The following are sampling techniques we will use in this thesis:

- **SMOTE(synthetic minority oversampling technique)** is a oversampling method. It produces new minority class examples by interpolating between one randomly selected example and one of its k-nearest neighbors.[15]

- **Tomek links** is defined as follows: "Given two example Ei and Ej that belonging to different classes, and we use $d(Ei, Ej)$ to represent the distance between Ei and Ej. In the original design, euclidean distance is used to measure the distance between two examples. If there does not exist any El such that satisfy $(Ei, El) < d(Ei, Ej)$ or $d(Ej, El) < d(Ei, Ej)$, then we can call $A(Ei, Ej)$ pair a Tomek link." [16]

- **Edited Nearest Neighborhood(ENN)** [17] ENN works by finding the k-nearest neighbors of each observed example, and then checking if the majority class of its k-nearest neighbors is the same as the class of the observed example. An observed example and its K-nearest neighbors will be deleted if the majority class of the observed example's k-nearest neighbors is different from its class.

- **SMOTE + Tomek links** "Because interpolating minority class examples will expand the minority class clusters, causing synthetic examples that belonging to minority class entering the majority class space too deeply." [6] To create better-defined class clusters, Gustavo et al.[6] applied Tomek links to the oversampled training set as a data cleaning method. Therefore, not only deleting the majority class examples that make up Tomek links, but also examples from another class(minority class).

- **SMOTE + ENN** "Compared with Tomek links, ENN will delete more examples and hence provide a deeper data cleaning." [6] ENN is used to delete examples from both

classes. Therefore, any example that is misclassified by the majority of its k nearest neighbors is deleted from the dataset. Same as Tomek links, ENN will remove examples from both classes.

2.3 Lossless De-biasing with Over-sampling Techniques

In Zhou et al.[5] research, they proposed a lossless de-biasing technique to improve the fairness of machine learning. They first analyzed the dataset and divided it into several subgroups according to protected attributes. Then they use an oversampling technique: SMOTE to synthesize data for the underrepresented group. They formalize their debiasing problem as bounding the discrepancy between the distributions of the privileged group D_p and augmented unprivileged group D'_u.

Problem. In a binary classification task, given a training dataset with its domain features X, a probability distribution D over X, they use D_p and D_u represent the distributions over privileged and unprivileged group respectively.

$$d(D_p, D'_u) = 2 \sup_{b \in B} |Pr_{D_p}(b) - Pr_{D'_u}(b)|,$$

above is a standard measurement of variation divergence for two distributions, where D'_u is the distribution of unprivileged group after augmented, and B is a collection of measurable subsets under D_p and D'_u.

Their formalization of the problem is inspired by the concept of transfer learning: if two populations differ only in demographic and socioeconomic backgrounds, then within a limited prediction errors and positive rates difference, models trained on one population (e.g. white defendants) should be applicable to other populations (e.g. black defendants).

We can consider distributions over the privileged group and the augmented unprivileged group as source domain and target domain, and we want the classifier's target error can be close to its empirical source error. Ben-Divid et al.[18] have provided a bound on the target domain error of the classifier in terms of its source domain error and the divergence measure between the

two domains. How the divergence between two domains connect to debiasing will be discussed later.

After obtaining the original data, the first thing to do is data analysis. We usually distinguish between privileged and unprivileged groups based on our social experience and the phenomena observed in the dataset. In terms of dataset analysis, Zhou et al.[5] defined the majority and privileged group as below: Given a demographic attribute A = {a1, a2} in a dataset X, the majority group defined on A is $X_{A=a^*} \subset X$ if $Pr(A = a^*|X) > Pr(A \neq a^*|X)$ where $a^* \in \{a1, a2\}$, the privileged group defined on A is $X_{A=a^*} \subset X$ if, historically, $Pr(y = 1|X_{A=a^*}) > Pr(y = 1|X_{A\neq a^*})$ where $a^* \in \{a1, a2\}, y \in \{0, 1\}$ and y=1 is favored.

Taking the Adult Census dataset as an example, we can observe from historical data that the average income of male is generally higher than that of female. To be more specific, male will have a higher chance of obtaining a high-income result than female, which can also be derived from the previous definition to get the same insight.

2.3.1 Theoretical Justification

In Zhou et al.[5], they provided a theoretical justification to explain how to connect the relation between the debiasing and bounding variation divergence between two distributions estimated using samples from D_p and D'_u.

Recall the following theorem provided by Ben-David et al.,2010[18]:

Theorem 1 *For a hypothesis h,*

$$\epsilon_T(h) \leq \epsilon_S(h) + d_1(D_S, D_T) + \min(E_{D_S}[|f_S(x) - f_T(x)|], E_{D_T}[|f_S(x) - f_T(x)|]).$$

where $\epsilon_T(h)$ and $\epsilon_S(h)$ is target error and source error, last term is the difference in labeling functions across two domains.

To fit the debias problem here, they replace the terms in Theorem 1 as follows:

$$\epsilon_{D_p}(h) - \epsilon_{D_u}(h) \leq d(D_p, D_u) + \lambda \tag{2.7}$$

where $\lambda = min_{h \in H}[\epsilon D_p(h) + \epsilon D_u(h)]$. Therefore, the difference of predictive errors from D_p and D_u by hypothesis h can be bounded by $d(D_p, D_u)$, which is their variation divergence. However, here comes an issue that "variation divergence cannot be accurately estimated from limited samples and hence has limited usefulness."[19]

To address the issue mentioned above, Ben-David et al.[18] proposed another measurement called $\mathcal{H}-divergence$, which can be estimated from finite samples to measure divergence between two distributions.

Definition 2.3.1 (H-divergence) *[18] Given a domain X with two probability distributions D and D' over X, let H be a hypothesis class of finite VC dimension on X, I(h) represent the set such that* $x \in I(h) \Leftrightarrow h(x) = 1$. *The* $\mathcal{H}-divergence$ *between D and D' is:*

$$d_H(D, D') = 2 \sup_{h \in H} [Pr_D[I(h)] - Pr_{D'}[I(h)]] \tag{2.8}$$

With the $\mathcal{H}-divergence$, they make a little modification to get the formula below and therefore bound the discrepancy between favorable outcome by h from distribution D_p and D'_u by $\mathcal{H}-divergence$:

$$d_H(D_p, D'_u) = 2 \sup_{h \in H} [Pr_{D_p}[I(h)] - Pr_{D'_u}[I(h)]] \tag{2.9}$$

$$\geq 2|Pr_{D_p}[h(x) = 1] - Pr_{D'_u}[h(x) = 1]| \tag{2.10}$$

$$|Pr_{D_p}[h(x) = 1] - Pr_{D'_u}[h(x) = 1]| \leq \frac{1}{2} d_H(D_p, D'_u). \tag{2.11}$$

And then by using Lemma 1 of Ben-David et al.[18]:

$$d_H(D, D') \leq \hat{d_H}(U, U') + 4\sqrt{\frac{dlog(2m) + log(\frac{2}{\delta})}{m}} \tag{2.12}$$

where U and U' are samples of size m from the distributions D and D'. Given any $\delta \in (0, 1)$, with probability at least $1 - \delta$, as the sample size increases, the empirical $\mathcal{H}-divergence$ can asymptotically converge to the true $\mathcal{H}-divergence$. Combining the equation above, Zhou et

al.[5] provided two inequality as follows, and thus we can bound the difference of predictive errors and the discrepancy of favorable predictions on data U_p and U_u of size m from D_p and D_u with the concept of $\mathcal{H}-divergence$:

$$\Delta\epsilon(h(x)=1) \leq \frac{1}{2}\hat{d_H}(U_p, U_u) + 2\sqrt{\frac{dlog(2m)+log(\frac{2}{\delta})}{m}} \tag{2.13}$$

where $\Delta\epsilon(h(x)=1) = |Pr_{D_p}[h(x)=1] - Pr_{D_u}[h(x)=1]|$.

$$\Delta\epsilon(h) \leq \frac{1}{2}\hat{d_H}(U_p, U_u) + 2\sqrt{\frac{dlog(2m)+log(\frac{2}{\delta})}{m}} + \lambda \tag{2.14}$$

where $\Delta\epsilon(h) = \epsilon D_p(h) - \epsilon D_u(h)$ and $\lambda = min_{h\in H}[\epsilon D_p(h) + \epsilon D_u(h)]$.

As inequality (2.13) and (2.14) suggest, we can bound the difference of favorable predictions (left term in inequality 2.13) and the difference in predictive errors (left term in inequality 2.14) by making the distribution divergence of the two groups D_p and D_u smaller, especially when one group is underrepresented in terms of favorable prediction. We can think of the objective of the oversampling method we use in this thesis as minimizing $\hat{d_H}(U_p, U_u)$ in inequality (2.13) and (2.14).

We can now connect the relation between the left terms in inequality (2.13) (2.14) and fairness measurement indicators. Favorable predictions difference(left term in inequality 2.13) can correspond to the Statistical Parity Difference, both of which are to pursue the same chance of obtaining favorable results between different demographic groups. Predictive errors(left term in inequality 2.14) can correspond to Equal Opportunity Difference and Average Odds Difference, both taking into account the performance of the classifier. And what they both have in common is that we all want to minimize these differences. Therefore, when the difference between favorable predictions and predictive errors is bounded, correspondingly, a promising result in fairness metrics can be obtained.

Chapter 3

Proposed Methods

Previous research has explored how different sampling techniques including oversampling and undersampling can improve machine learning performance. Oversampling techniques tackled the class imbalanced problem, undersampling techniques further reduce the number of the noisy or borderline examples and hence solve class overlapping issues. On the other hand, the use of over-sampling technique SMOTE for data pre-processing has been proposed by previous studies and has theoretical justification and experimental results show that it can effectively remove the bias in machine learning and improve fairness. However, research on undersampling techniques for de-bias in machine learning still remains unexplored.

In addition, in the field of machine learning algorithm debiasing research, there exists a trade-off between fairness and accuracy, that is, the machine learning model may suffer from a drop in accuracy while debiasing. In the research of the COMPAS algorithm, we can also find that they encounter the same issue. Corbett-Davies, Goel, Pierson, Huq, and Feller point out "an inherent tension between minimizing violent crime and satisfying common notions of fairness." [20] In their conclusion, they argue that maximizing public safety could result in unfair decisions for defendants of color. However, if social justice is to be met, it may lead to the release of high-risk defendant groups, which will negatively affect public safety.

A more intuitive understanding of the trade-off between fairness and accuracy is that in the classification task, what accuracy measures is focusing on the overall data, while the fairness is measured after the data is grouped and focused on the difference in performance between these groups. These two objectives are different in nature. When we use the pre-processing method, we change the structure and composition of the training set data, which may affect the distribution of a certain group or the whole group, making the model trained on dataset after processing may not be able to generalize effectively to test data, which also has some chances to lead to a decrease in classification performance.

Using undersampling methods such as ENN or Tomek links can effectively reduce the num-

ber of instances in the data sample space that are regarded as noisy or borderline, thereby reducing misclassification and obtained a better performance. Another advantage of undersampling worth mentioning is that it can reduce the risk of machine learning algorithms being biased towards the majority group.

Therefore, combining the characteristics of undersampling techniques mentioned above, instead of using SMOTE alone, we propose to use two combinations of sampling techniques, which are SMOTE+Tomek links and SMOTE+ENN, as our method on data pre-processing for debias. We believe that performing undersampling on the oversampled data can further enhance the de-biasing effect and maintain the accuracy performance in the mean time. Take the Adult dataset for example, we use SMOTE to oversample the underrepresented group(the unprivileged favored group), to make the difference between the distributions of the privileged group and the unprivileged group closer, then we use ENN or Tomek links to undersample the unprivileged group. The datasets after undersampling may not maintain a predetermined distribution among their subgroups. In order to conform to Zhous' theoretical justification, we also make a little modification on the TomekLinks algorithm, and the detailed modification process will be explained in the next section.

After all sampling techniques are implemented on the dataset, we proceed to experiment in two baseline classifiers: Logistic Regression and Random Forest. Finally, we observe the experimental results obtained above and compare them with the results of using SMOTE alone.

3.1 Modified Tomek Links

Since Tomek links as an undersampling method, when it deletes samples for a group, it will first find a set of Tomek pairs and delete them, but the design of the original algorithm will delete two belonging to different class at the same time during the deletion process, which made it impossible to maintain our predetermined ratio, so we modified the way Tomek links were deleted.

Referring to the formula guarantee provided by Zhou's, we know that as long as the first and second terms on the right side of the inequality 2.13 and the inequality 2.14 are minimized, the fairness can be improved (ex: minimizing Statistical Parity Difference). So the motivation

for modifying the algorithm to maintain the same proportion of the two groups is that we hope to control and minimize the first term on the right side of the inequality 2.13 and the inequality 2.14, which is the empirical H-divergence between the two groups.

Compared to the original Tomek links, we take the resulting distribution ratio into account as one of our parameters. After finding Tomek pairs, delete all samples belonging to the majority group in the pairs first, then sort the samples belonging to the minority group in the pair in ascending order of distance, and delete the samples in the minority group according to the proportion. We can see the modified part of Algorithm 3.1. The previous part of Algorithm 3.1 is the same as the original Tomek links as undersampling. Starting from line 15, it is our modified part of the algorithm. Line 15 sorts the samples ascendingly belonging to the Tomek pairs in the minority group according to the distance from their nearest neighbor, and line 17 extract samples according to the group distribution ratio parameter. The reason why Tomek pairs are sorted in ascending order of distance is because we know that Tomek links will treat the samples that makeup Tomek pairs as noisy or borderline examples, so the closer two samples in a Tomek pair are, the greater chance they will interfere with the performance of the machine learning model. So when we delete samples belonging to the minority group in Tomek pairs, the samples with smaller distances from their nearest neighbor are preferentially deleted.

Algorithm 3.1 Modified Tomek links undersampling

Input: D : Datasets to resample, R : Group distribution ratio , $Mlabel$: majority label ,$mlabel$: minority label

Output: D' : Resampled Datasets

$TLMajList \leftarrow [\,]$
$TLminList \leftarrow [\,]$
for $data$ in D **do**
    if $data.y$ not equals $data.NN.y$ **then** ▷ NN stands for nearest neighbor
        if $data.NN$ equals $data$ **then**
            if $data.y$ equals $Mlabel$ **then**
                $TLMajList$.push([$data.id$,$data.distance$]) ▷ data.distance: distance between data and data.NN
            else if $data.y$ equals $mlabel$ **then**
                $TLminList$.push([$data.id$,$data.distance$])
            end if
        end if
    end if
end for

Sort $TLminList$ by distance with ascending order.

$TLminList \leftarrow TLminList[0 \ldots int(TLMajList.length \times R)]$

for $TLdata$ in $TLmajList$ and $TLminList$ **do**
    delete $TLdata$ from D
end for

$D' \leftarrow D$
return D'

Chapter 4

Experimental Results

In this section, we will discuss the datasets we experiment with in this thesis, and how we implement the sampling method to mitigate the bias in the dataset, finally we will analyze the result using performance metrics and fairness metrics. We use AI fairness 360, a toolkit provided by IBM [21], and Imblearn [22] library as our tool for the experiment. We use two baseline classification algorithms: Logistic Regression and Random Forest on our dataset, the experimental dataset is divided into training set (70%) and test set (30%), and according to the prediction result, observe the change of the fairness metric.

4.1 Datasets

In this thesis, we use the following open datasets provided by UCI[23]: Adult, German Credit, Bank dataset and ProPublica: COMPAS dataset for our experiment, and the detail information of the dataset such as sample size and attributes being used will introduced in appendix:

- **Adult Census**

  Adult Census dataset was donated to UCI Machine Learning Repository in 1994. Its object is to predict whether a person's income can achieve 50K a year. The Adult Census dataset has 48,842 entries, and 45,222 after instances with unknown values are removed. In this dataset, we use all attributes except 'fnlwgt'. In the Adult Census dataset, 'Male' in sex is the privileged group, and the favored outcome is '>50K'.

- **COMPAS**

  COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions. It's a popular business analysis used by judges and parole officers to score criminal defendants' likelihood of reoffending (recidivism). The COMPAS dataset provided by ProPublica has 5278 records. We use 'sex', 'race', 'age', 'priors counts', 'charge degree'

these five attributes in the experiment. In COMPAS, race is the protected attribute, the privileged group is 'Caucasian' and the unprivileged group is 'African-American', and the favored class is 'no recidivism'.

- **German Credit**

  German Credit dataset contains 1,000 entries. Each entry(person) is classified as good or bad credit risks according to a set of attributes. We use the following attributes in our experiment : 'age', 'sex, 'duration', 'purpose', 'credit amount', 'housing', 'job', 'num dependents', 'saving status'. The privileged group is 'Old' and the unprivileged group is 'young'. The favored outcome is 'Good Credit'.

- **Bank**

  Bank Marketing dataset is also contributed in UCI ML Repository. It's related to a marketing campaign of a banking institution and its classification task is to predict whether a client will subscribe to a term deposit. The privileged group is 'aged'(age>=25) and the favored outcome is 'yes'(subscribe a deposit). In Bank dataset, the privileged group accounts for a huge majority of the overall population.

The datasets we experiment with contain two cases. in terms of obtaining the favored outcome, the privileged group is the majority and the unprivileged group is the majority. We can use the definitions mentioned in Chapter 2.3 to help us analyze datasets and to find out which group is belonging to privileged/unprivileged and favored/unfavored group.

4.2 Basic Random Sampling Techniques

We first use the basic sampling method: random over and undersampling to experiment on the dataset. Random over and undersampling are: randomly selecting samples from the minority group and duplicating them, and randomly selecting samples from the majority group and deleting them respectively.

We use these two basic sampling methods to try to confirm that the previous theoretical justification is feasible in experiments. From the results in Tables 4.1 and 4.2, it can be seen

Logistic Regression						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Orig.	0.74257	-0.17696	0.28292	-0.10133	-0.12433
	Rand. Over	0.74147	-0.0878	0.62004	**0.06697**	0.15255
	Rand. Under	0.74659	**-0.04803**	**0.80283**	0.11177	0.19883
Compas	Orig.	**0.66774**	-0.29995	0.60556	-0.27106	-0.20181
	Rand. Over	0.66127	**-0.06382**	**0.88358**	**-0.029**	**0.01209**
	Rand. Under	0.66649	-0.12177	0.78783	-0.08833	-0.03416
German	Orig.	0.55875	-0.12083	0.86524	-0.09866	-0.12378
	Rand. Over	**0.57107**	-0.03462	0.96138	**-0.00476**	-0.03864
	Rand. Under	0.55022	**0.00056**	**1.00062**	0.02406	**-0.01219**
Bank	Orig.	**0.58159**	0.02494	1.82188	0.00588	0.01334
	Rand. Over	0.5799	0.02066	1.69316	-0.00204	-0.00276
	Rand. Under	0.58111	**0.02036**	**1.67606**	-0.00333	-0.00535

Table 4.1: Basic sampling method in Logistic Regression

Random forest						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Orig.	0.74257	-0.17696	0.28292	-0.10133	-0.12433
	Rand. Over	0.74147	-0.0878	0.62004	**0.06697**	0.15255
	Rand. Under	0.74659	**-0.04803**	**0.80283**	0.11177	0.19883
Compas	Orig.	**0.66774**	-0.29995	0.60556	-0.27106	-0.20181
	Rand. Over	0.66127	**-0.06382**	**0.88358**	**-0.029**	**0.01209**
	Rand. Under	0.66649	-0.12177	0.78783	-0.08833	-0.03416
German	Orig.	0.55875	-0.12083	0.86524	-0.09866	-0.12378
	Rand. Over	**0.57107**	-0.03462	0.96138	**-0.00476**	-0.03864
	Rand. Under	0.55022	**0.00056**	**1.00062**	0.02406	**-0.01219**
Bank	Orig.	**0.58159**	0.02494	1.82188	0.00588	0.01334
	Rand. Over	0.5799	0.02066	1.69316	-0.00204	-0.00276
	Rand. Under	0.58111	**0.02036**	**1.67606**	-0.00333	-0.00535

Table 4.2: Basic sampling method in Random forest

that random oversampling and random undersampling used in Logistic Regression or Random Forest in the four datasets can effectively increase the disparate impact and reduce the statistical parity difference. It represents that such operations on the distribution of data can effectively obtain a fairer outcome.

However, the operation of random oversampling on the dataset may cause overfitting, and random undersampling may inevitably discard useful information in the data.[6]

ADULT					
	SMTL	SMTL (Correlation)		SMTL	SMTL (Correlation)
Balanced Acc.	0.76099	0.74885	Balanced Acc.	0.76796	0.7683
AUC	0.87139	0.86785	AUC	0.88844	0.88797
DI	0.31102	0.72507	DI	0.33103	0.32271
AOD	-0.09985	0.0964	AOD	-0.07998	-0.0854
AOE	-0.09985	0.0964	AOE	0.07998	0.0854
SPD	-0.20857	-0.06691	SPD	-0.18573	-0.18863
EOD	-0.08395	0.18849	EOD	-0.07407	-0.08312

GERMAN					
	SMTL	SMTL (Correlation)		SMTL	SMTL (Correlation)
Balanced Acc.	0.61081	0.57353	Balanced Acc.	0.56698	0.59211
AUC	0.69321	0.68361	AUC	0.65632	0.67191
DI	0.85344	0.98061	DI	0.93644	0.90039
AOD	-0.06033	0.01136	AOD	-0.03162	-0.05385
AOE	0.03386	0.01773	AOE	0.04697	0.05385
SPD	-0.10658	-0.01738	SPD	-0.05163	-0.08401
EOD	-0.03864	-0.00638	EOD	-0.05405	-0.04611

COMPAS					
	SMTL	SMTL (Correlation)		SMTL	SMTL (Correlation)
Balanced Acc.	0.66127	0.66127	Balanced Acc.	0.64986	0.64979
AUC	0.70732	0.70999	AUC	0.70674	0.70575
DI	0.88358	0.88358	DI	0.94871	0.95307
AOD	-0.029	-0.029	AOD	0.00275	0.00573
AOE	0.0411	0.0411	AOE	0.05007	0.04758
SPD	-0.06382	-0.06382	SPD	-0.02843	-0.02582
EOD	0.01209	-0.06383	EOD	0.05283	0.05332

BANK					
	SMTL	SMTL (Correlation)		SMTL	SMTL (Correlation)
Balanced Acc.	0.57942	0.58242	Balanced Acc.	0.59054	0.5904
AUC	0.72104	0.71962	AUC	0.69747	0.69638
DI	1.71504	1.63478	DI	2.25357	2.23123
AOD	-0.00126	-0.00517	AOD	0.03324	0.03098
AOE	0.00126	0.00517	AOE	0.03324	0.03098
SPD	0.02104	0.0196	SPD	0.05725	0.05744
EOD	-0.00146	0.01961	EOD	0.03287	0.05744

Table 4.3: TomekLinks Distance Compararison

4.3 Combining Sampling Techniques for Debias

4.3.1 TomekLinks Distance Measurement Discussion

In the design of Tomek Links algorithm, the distance function to find the nearest neighbor is an optional variable. The original design was to use the Euclidean distance as a measure of the distance between two points. We also tried using correlation for the measurement. Because compared with Euclidean distance, correlation can be well applied to the situation when the data scale is inconsistent, and it also solves the problem of grade inflation. The measurement and comparison results of two different distances are placed in table 4.3. But in our experiments, our dataset was normalized before feeding into the model. Therefore, in the experimental results table 4.3, we cannot see a significant difference between the two measurement methods. So we finally chose to use the original design, which is to use the Euclidean distance.

4.3.2 Numerical Results

Table 4.4 (Logistic Regression) and Table 4.5 (Random Forest) are the result of the classification performance and fairness metric of each dataset. We first look at Statistical Parity Difference and Disparate Impact, except for the COMPAS dataset in LR slightly inferior to the SMOTE method one, in the two classifier models the SMOTE+ENN we use can achieve the fairest result in the experiments of each dataset, especially the German dataset after sampled by SMOTE+ENN, the Disparate Impact is improved to almost close to 1.0 (1.00062/0.99936), and the Statistical Parity Difference is also reduced to approximately 0.0 (0.00056/-0.00076),

which can be regarded as the fairest outcome of all. The Adult dataset in LR has a significant improvement (from 0.28292 to 0.80283) and the Bank dataset in RF, obtained an obviously drop to close to 1.0(from 2.12165 to 1.15362) in Disparate Impact.

Then we observed Average Odds Differences, despite that only the Bank dataset in LR achieves the fairest result after preprocessing of the combination of sampling techniques, in Random Forrest we can use SMOTE+ENN or SMOTE+Tomek links to make the four datasets get the most promising outcome. In terms of Equality Odds Differences, we can conclude that in all experimental data in LR and in RF, the best outcome is obtained by SMOTE+ENN or SMOTE+Tomek links, except that the data of COMPAS dataset in LR does not match.

It is worth mentioning that the method of combination sampling technique is just like the oversampling method used in the previous study, which can also improve the fairness metric while maintaining classification utility(balanced accuracy) well.

In the experimental results of Modified Tomek links in Logistic Regression (Table 4.4), we can find that compared with the original dataset, SMOTE + Modified Tomek links have also improved in the group fairness metric, which once again echoes the correlation between the distribution ratios of different groups in the training set and the group fairness.

The reason why Modified Tomek links is not done for the COMPAS dataset is that only one set of TomekPairs is found in it, so we think that Modified Tomek links have no significant difference in results on the COMPAS dataset.

4.3.3 Discussions

SMOTE+ENN obtained the fairest results (SPD / DI) in the Adult, German, and Bank dataset under two classification models. Figure 4.1 shows the distribution ratio of the dataset after different sampling techniques. See Adult, German, and the Bank dataset, after the sampling operation of SMOTE+ENN, the proportion of originally underrepresented groups (ex: Adult: Unpri-favored and Bank: Unpri-unfavored) has been lifted up a little more compared with other sampling techniques.

Figure 4.2 shows the positive rate between privileged and unprivileged group of each dataset. We still focus on the results of SMOTE+ENN, because they achieve fairer performance in both

fairness indicators SPD and DI. In the Adult and German datasets, because the target we oversampled in these two datasets is the unprivileged-favored group, we can observe that the unprivileged positve rate is significantly lift up under two classifier models, which is closer to the positive rate of the privileged group. In addition, in the Bank dataset, we oversampled the unprivileged-unfavored group, and the positive rate of the unprivileged group in RF therefore dropped significantly and approached the opposite group.

We think it is because ENN has the characteristic of more aggressively deleting the target sample's neighbor samples, it suppresses the proportion of the opposite groups of the original underrepresented group (ex: Adult: Unpri-unfavored and Bank: Unpri-favored), so SMOTE+ENN will make the target group get a higher scale boost.

On the other hand, we also observed that although Modified Tomek Links can perfectly control our predetermined groups proportion and achieve the goal of minimizing the first term (divergence) on the right-hand side of inequality (2.13) and inequality (2.14), it is unable to optimize fairness in all experiments. In addition, although the Smote+ENN (or Smote+Tomeklinks) in this experiment cannot control the final groups ratio and the final sample size m in the implementation process, it sometimes achieves a better performance in the experiment. From the perspective of inequality (2.13), even if we control the first term on the right to minimum, it will still be affected by the second term, so we think that theoretically, as long as the first and second terms on the right side of inequality (2.13) can be controlled and optimized at the same time, we can guarantee to obtain the best fairness (SPD) result.

Logistic Regression						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Orig.	0.74257	-0.17696	0.28292	-0.10133	-0.12433
	Smote	0.74147	-0.0878	0.62004	**0.06697**	0.15255
	Smote+ENN	0.74659	**-0.04803**	**0.80283**	0.11177	0.19883
	Smote+TL	**0.76099**	-0.20857	0.31102	-0.09985	**-0.08395**
	Smote+TL*	0.74583	-0.06748	0.7165	0.09261	0.18189
COMPAS	Orig.	**0.66774**	-0.29995	0.60556	-0.27106	-0.20181
	Smote	0.66127	**-0.06382**	**0.88358**	**-0.029**	**0.01209**
	Smote+ENN	0.66649	-0.12177	0.78783	-0.08833	-0.03416
	Smote+TL	0.66127	-0.06382	0.88358	-0.029	0.01209
	Smote+TL*	-	-	-	-	-
German	Orig.	0.55875	-0.12083	0.86524	-0.09866	-0.12378
	Smote	**0.57107**	-0.03462	0.96138	**-0.00476**	-0.03864
	Smote+ENN	0.55022	**0.00056**	**1.00062**	0.02406	**-0.01219**
	Smote+TL	0.57107	-0.03462	0.96138	-0.00476	-0.03864
	Smote+TL*	0.56861	-0.03049	0.96583	-0.00186	-0.03282
Bank	Orig.	**0.58159**	0.02494	1.82188	0.00588	0.01334
	Smote	0.5799	0.02066	1.69316	-0.00204	-0.00276
	Smote+ENN	0.58111	**0.02036**	**1.67606**	-0.00333	-0.00535
	Smote+TL	0.57942	0.02104	1.71504	**-0.00126**	**-0.00146**
	Smote+TL*	0.58111	0.02036	1.67606	-0.00333	-0.00535

Table 4.4: Comparision of different sampling method in Logistic Regression

Random Forest						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Orig.	0.76907	-0.19044	0.31569	-0.08641	-0.08551
	Smote	0.76832	-0.18615	0.32926	-0.08086	-0.07579
	Smote+ENN	**0.7755**	**-0.15185**	**0.45715**	**-0.00049**	**0.06433**
	Smote+TL	0.76796	-0.18573	0.33103	-0.07998	-0.07407
	Smote+TL*	0.76831	-0.18663	0.32813	-0.08196	-0.07775
COMPAS	Orig.	**0.6581**	-0.24742	0.67884	-0.2194	-0.15587
	Smote	0.64904	-0.02802	0.9493	0.00292	0.05326
	Smote+ENN	0.63489	**-0.00846**	**0.98264**	-0.03184	**0.05099**
	Smote+TL	0.64986	-0.02843	0.94871	**0.00275**	0.05283
	Smote+TL*	-	-	-	-	-
German	Orig.	**0.59577**	-0.10856	0.87115	-0.07688	-0.07425
	Smote	0.59557	-0.08857	0.89404	-0.05413	-0.06753
	Smote+ENN	0.58435	**-0.00076**	**0.99936**	**0.03142**	**0.0193**
	Smote+TL	0.59476	-0.08036	0.90391	-0.04697	-0.05405
	Smote+TL*	0.59467	-0.09173	0.89111	-0.05919	-0.06251
Bank	Orig.	0.58848	0.09021	2.12165	0.07176	0.0836
	Smote	**0.59058**	0.06904	2.48286	0.04918	0.0596
	Smote+ENN	0.58695	**0.00711**	**1.15362**	-0.03367	-0.06637
	Smote+TL	0.59054	0.05725	2.25357	**0.03324**	**0.03287**
	Smote+TL*	0.59059	0.07755	2.66086	0.05839	0.07014

Table 4.5: Comparision of different sampling method in Random Forest

Consistency Score				
	Original	Smote	SM+ENN	SM+TomekLinks
Adult	0.84792	0.85363	**0.87178**	0.85367
COMPAS	0.60514	0.59378	**0.72738**	0.59407
German	0.70171	0.73107	**0.77134**	0.74013
Bank	0.85875	0.86325	**0.86812**	0.86553

Table 4.6: Individual fairness metric for each sampling technique

4.3.4 Individual Fairness Measurement

In addition to group fairness, we also found that our proposed method also performs well in individual fairness. We calculate the consistency score for the sampled dataset. In Table 4.6, we can find once again that SMOTE+ENN, a combination sampling technique achieves the best score in all four datasets. In addition, although another combination sampling technique, SMOTE+Tomek links, is slightly inferior in the COMPAS dataset, it also performs better on the other three datasets than the original and the SMOTE one.

We believe that undersampling plays a role here, because ENN and Tomek links, as undersampling techniques, can delete examples that are considered noisy or borderline. And these examples that are considered noisy or borderline are usually instances that are close to each other in the sample space but do not belong to the same class.

Compared to our method, examples of noisy or borderline still exist in the data that only use the SMOTE oversampling technique and have not been processed by ENN or Tomek links. ENN and Tomek links create better-defined class clusters after deleting those examples, and also reduce the average difference between the prediction of each example and the prediction results of its neighboring example points in equation 2.1, which eventually leads to that the overall consistency score can be effectively lift up.

4.4 Additional Result and Comparison

4.4.1 Experimental Result of Another Ensemble Model: XGBoost

We also use an another type of ensemble learning model: XGBoost for our experiments, and the results are presented in table 4.8. We can find that all sampling techniques can improve

Logistic Regression						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Opt-Prepro	0.7013	-0.0722	0.7895	**-0.0487**	-0.0429
	LFR	0.7119	-0.0721	**0.8035**	0.0132	**0.0186**
	Smote+ENN	**0.74659**	**-0.04803**	0.80283	0.11177	0.19883
COMPAS	Opt-Prepro	**0.6752**	**-0.097**	**0.8534**	**-0.0678**	-0.0691
	LFR	0.6667	-0.1863	0.7562	-0.1468	-0.1105
	Smote+ENN	0.66649	-0.12177	0.78783	-0.08833	**-0.03416**

Table 4.7: Comparison with other preprocessing techniques

XGBoost						
	Method	Balanced Acc.	SPD	DI	AOD	EOD
Adult	Orig.	0.79664	-0.18648	0.31265	-0.08126	-0.09309
	Smote	0.80034	-0.17786	0.34994	0.05277	**-0.03877**
	Smote+ENN	**0.80204**	**-0.14384**	**0.47488**	**0.00351**	0.04629
	Smote+TL	0.79763	-0.17359	0.35657	-0.05521	-0.04955
COMPAS	Orig.	**0.66181**	-0.23663	0.69012	-0.20571	-0.15501
	Smote	0.64905	-0.03548	0.93621	-0.00333	**0.03751**
	Smote+ENN	0.632	**0.01057**	**1.02068**	0.03338	0.11515
	Smote+TL	0.65001	-0.02924	0.94743	**0.00294**	0.04599
German	Orig.	**0.59781**	-0.16585	0.77946	-0.12937	-0.16297
	Smote	0.59219	-0.15203	0.80324	-0.12435	-0.11009
	Smote+ENN	0.58435	**0.02792**	**1.03734**	**0.05627**	**0.07445**
	Smote+TL	0.58166	-0.12995	0.83448	-0.10027	-0.12172
Bank	Orig.	0.5799	0.0225	1.68661	**0.00144**	**0.004**
	Smote	0.58451	0.03697	2.05232	0.00304	-0.01766
	Smote+ENN	**0.58531**	**0.00615**	**1.16578**	-0.03862	-0.08021
	Smote+TL	0.58348	0.00969	1.26957	-0.02438	-0.0457

Table 4.8: Comparision of different sampling method in XGBoost

the fairness measurement indicator very well, and the results are similar to the results of Random Forest, especially in SPD and DI, the combination of Smote+ENN achieves the best. The motivation we experimented with more types of models is to exclude the influence of machine learning models on the results. We do not expect the effectiveness of sampling techniques to improve fairness to vary depending on the machine learning model we choose.

4.4.2 Comparison With Fair Representation

In addition to the comparison between each sampling technique, we also make an additional comparison between the sampling technique and fair representation pre-processing techniques. In Table 4.7, we compare the experimental results of SMOTE + ENN with Optimized-

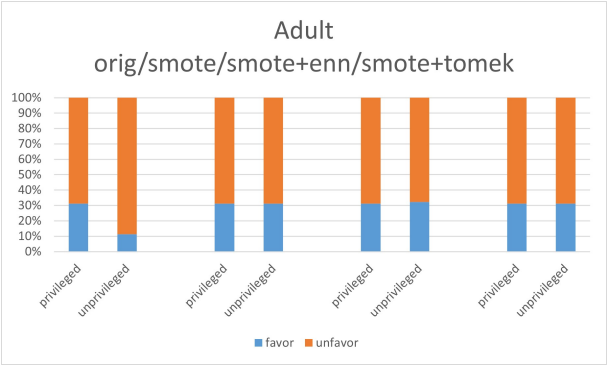
Preprocessing and Learning Fair Representation. SMOTE + ENN obtain better results in Adult: Balanced Accuracy(0.74659) and SPD(-0.04803), and in COMPAS: EOD(0.03416). Although the performance of improving fairness is about the same, we consider that sampling techniques are more intuitive and easier to understand and operate than the transformation of fair representation technique.

4.5 Shap Values Evaluations

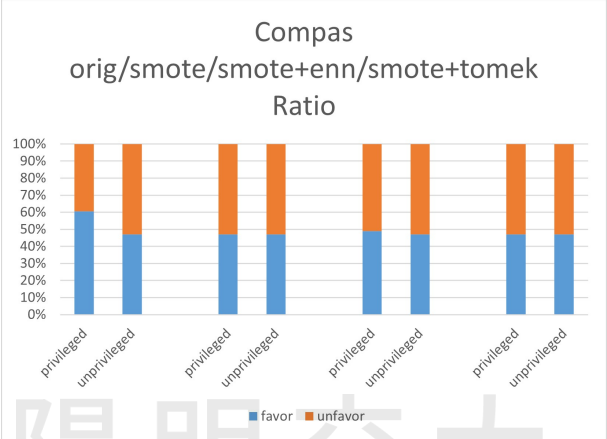
Finally, we evaluate SHAP values for each model trained on the sampled dataset. SHAP, which stands for SHapley Additive exPlanations, is an important and common tool in the field of explainable AI. SHAP values are derived from cooperative game theory and are used to interpret the prediction results of the model into the contribution of each factor. By calculating the SHAP values of each feature, the contribution of each feature to the prediction is evaluated.

In the design concept of this paper, we hope to improve fairness through sampling techniques. As mentioned earlier, while improving fairness, we hope that the mutual information between protected attributes and the model predictions should be as small as possible. So here we intuitively think and expect that sampling techniques can reduce the impact of protected attributes on the model prediction results.

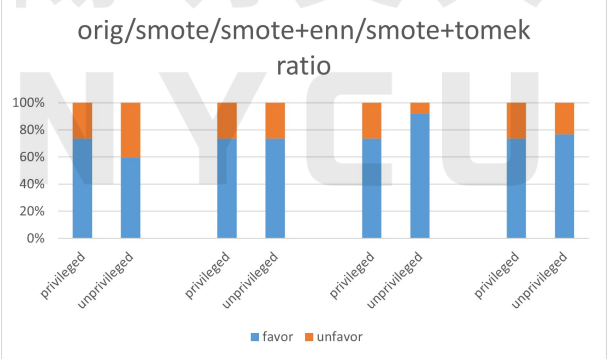
Figure 4.3 shows the evaluation results of SHAP values of the two datasets: Adult and German Credit's SHAP values. These two datasets are the datasets we believe that the sampling technique has the most outstanding performance in improving fairness. We can observe that in these two datasets, after the adjustment of sampling techniques, the influence of protected attributes(Adult: sex, German Credits: age) on the prediction results by the model show a decreasing trend, which to some extent echoes what we expected earlier, that is, we want to reduce the protected attribute's influence on model predictions. More precisely, we don't want our model predictions to be affected by protected attributes.



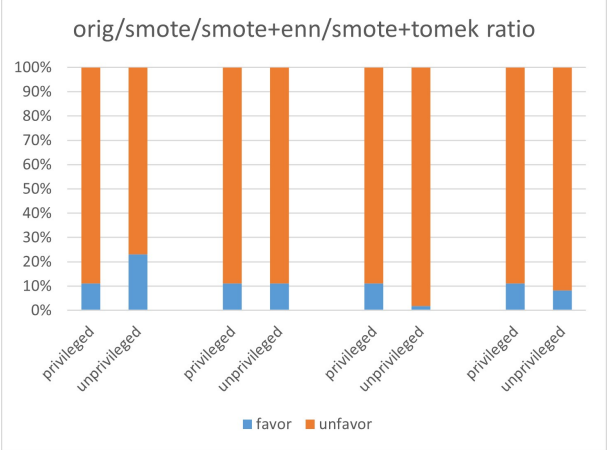
(a) Adult dataset



(b) COMPAS dataset



(c) German dataset



(d) Bank dataset

Figure 4.1: The ratio between privileged and unprivileged group

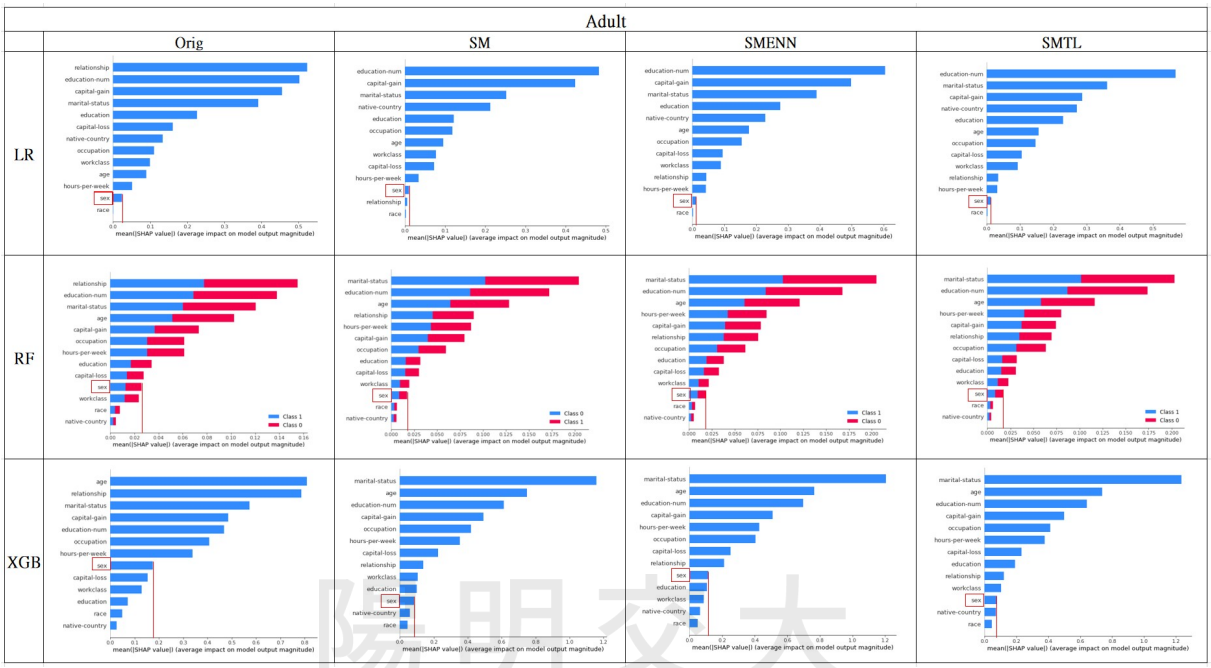
	Adult		COMPAS		German		Bank	
	Unprileged	Privileged	Unprileged	Privileged	Unprileged	Privileged	Unprileged	Privileged
Original	0.06982	0.24678	0.46049	0.76045	0.77586	0.89669	0.05528	0.03034
Smote	0.14327	0.23107	0.4844	0.54823	0.86206	0.89669	0.05048	0.02981
Smote+ENN	0.19558	0.24362	0.45218	0.57395	0.91379	0.91322	0.05048	0.03011
Smote+Tomek	0.09415	0.30272	0.4844	0.54823	0.86206	0.89669	0.05048	0.02943

(a) Positive rate between pri/unprivileged group in LR

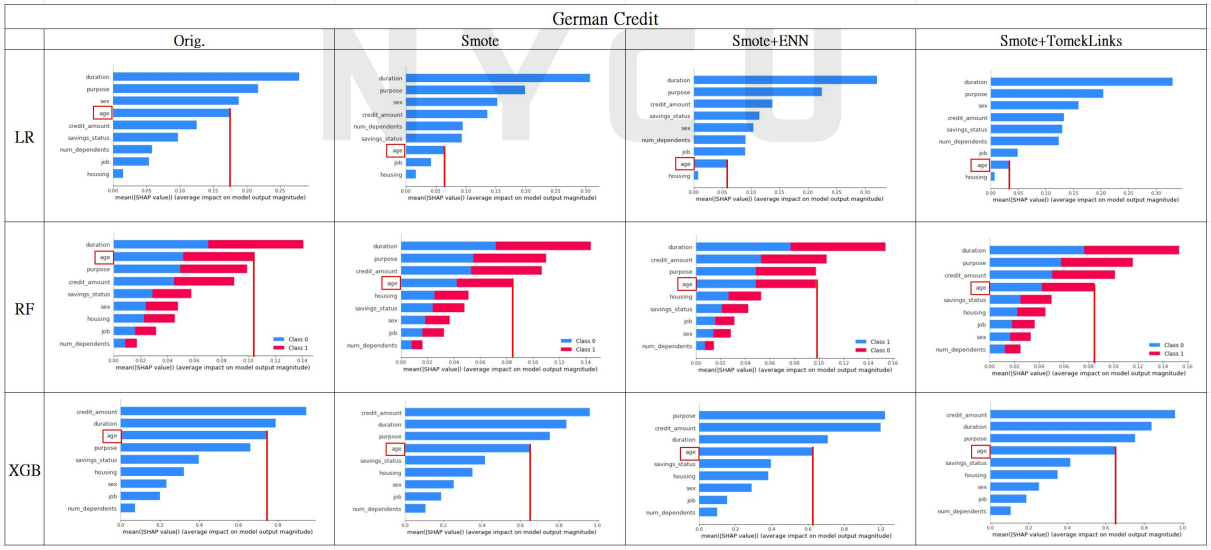
	Adult		COMPAS		German		Bank	
	Unprileged	Privileged	Unprileged	Privileged	Unprileged	Privileged	Unprileged	Privileged
Original	0.08794	0.27769	0.52234	0.76977	0.73275	0.84132	0.17067	0.08045
Smote	0.09138	0.27753	0.52068	0.54871	0.74655	0.83512	0.11562	0.04657
Smote+ENN	0.12788	0.27973	0.47931	0.48778	0.84137	0.84214	0.0536	0.04649
Smote+Tomek	0.0919	0.27763	0.52525	0.55369	0.75517	0.83553	0.10528	0.04672

(b) Positive rate between pri/unprivileged group in RF

Figure 4.2: Positive between pri/unprivileged group



(a) Shap values result on Adult Census



(b) Shap values result on German Credit

Figure 4.3: Shap Values Evaluations

Chapter 5

Conclusions and Future Work

5.1 Conclustions

In previous studies, the field of data pre-processing using the oversampling technique SMOTE for de-biasing in ML has been shown to be effective. However, the effect of undersampling on debiasing has not yet been studied. Therefore, in this thesis, we propose applying combination sampling techniques for debias in machine learning.

First, we use basic sampling techniques to confirm the relationship between fairness and groups distribution. Then, compared with the previous method of only using the oversampling technique(SMOTE), we obtained the fact that the combination sampling techniques (SMOTE+TL and SMOTE+ENN) achieve better results in both group fairness and individual fairness indicators while maintaining classification performance well through experimental results. We also discussed an interpretation that the improvement on fairness, in addition to the improvement in oversampling for underrepresented group, is largely due to the influence of undersampling in it: promoting underrepresented groups while suppressing the opposite groups to explain the result.

On the other hand, we modified the Tomek links algorithm to emphasize the relationship between the group ratio and the group fairness through experimental comparison results. We also experimented with another ensemble learning model: XGBoost, and the results are similar to the performance of the previous two baseline models. This removes the influence of machine learning model selection on the fairness improvement effect of sampling techniques. And we also compared SMOTE+ENN with other pre-processing methods and obtained the result that although it is not the best in fairness, we believe that the sampling technique is more intuitive and easy to use. Finally, through the analysis of SHAP values, we can also observe that sampling techniques have a positive effect on reducing the impact of protected attributes on model predictions.

In conclusion, we provide a new method in the field of pre-processing to improve the effect of de-biasing and conducted extensive experiments with different sampling techniques under different models which can be referenced by future researchers, model developers or even policy makers.

5.2 Future Work

In this thesis we experiment only on four well-known datasets, and in the future we plan to try more different datasets (ex: Medica Expenditure dataset). In this experiment, the combination sampling technique is based on the design in [6], that is, the target dataset is oversampled before being undersampled. We can try to change their order in the future to observe the effect.

In addition, in this experimental design, our combination sampling technology refer to the design in [5], which will only work on one demographic group at a time, so we also plan to redesign it to work on the overall datasets in the future. Finally, we also plan to explore the possibility of more different combination sampling techniques that combine other advanced sampling techniques and their effect on de-biaing in Machine Learning.

References

[1] N. T. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," *Brookings Institute: Washington, DC, USA*, 2019.

[2] J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang *et al.*, "Ai recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health*, 2022.

[3] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 358—364. [Online]. Available: https://doi.org/10.1145/3375627.3375865

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[5] Y. Zhou, M. Kantarcioglu, and C. Clifton, "Improving fairness of ai systems with lossless de-biasing," *arXiv preprint arXiv:2105.04534*, 2021.

[6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 20—29, jun 2004. [Online]. Available: https://doi.org/10.1145/1007730.1007735

[7] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3995–4004.

[8] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[10] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[11] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[12] F. Kamiran, S. Mansha, A. Karim, and X. Zhang, "Exploiting reject option in classification for social discrimination control," *Information Sciences*, vol. 425, pp. 18–33, 2018.

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[14] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," p. 561802, 2021.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[16] I. Tomek, "Two modifications of cnn," *IEEE Trans. Systems, Man and Cybernetics*, vol. 6, pp. 769–772, 1976.

[17] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408–421, 1972.

[18] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[19] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, 2000, pp. 259–269.

[20] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[21] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[22] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[23] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

Appendix A

Datasets

The following is the detailed information of the dataset used in the experiments of this thesis:

A.1 Datasets Information

	Adult Census	**COMPAS**	**German Credit**	**Bank**
Prediction Task	Predict whether a person's income can achieve 50K a year.	Score criminal defendants' likelihood of reoffending (recidivism).	Classify each person's credit risk good or bad.	Predict if the client will subscribe a term deposit.
# Of Entities	45,222	5,278	1,000	45,211
# Of Features For Training	13	5	9	10
Sensitive Attribute	Sex	Race	Age	Age

Table A.1: datasets details

Adult Census			
	Name	**Type**	**Values**
Attributes	age	Continuous	17∼90
	workclass	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
	education	Categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
	education-num	Continuous	1∼16
	marital-status	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
	occupation	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
	relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
	race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
	sex	Binary	Female, Male.
	capital-gain	Continuous	0∼99999
	capital-loss	Continuous	0∼4356
	hours-per-week	Continuous	1∼99
	native-country	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
Label	income	Binary	>50K, <=50K

COMPAS			
	Name	**Type**	**Values**
Attributes	sex	Binary	Female, Male.
	race	Categorical	Caucasian, African-American
	age_cat	Categorical	Less than 25 , 25 - 45, Greater than 45
	priors_count	Continuous	0∼38
	c_charge_degree	Binary	F, M
Label	Recidivated	Binary	Recidivated, Survived

German Credit			
	Name	**Type**	**Values**
Attributes	age	Continuous	19∼75
	duration	Continuous	4∼72
	purpose	Categorical	new car, used car, furniture/equipment, radio/tv, domestic appliances, repairs, education, retraining, business, others
	credit_amount	Continuous	250∼18424
	housing	Categorical	rent, own, for free
	job	Categorical	unemployed/ unskilled - non-resident, unskilled - resident, skilled employee / official, management/ self-employed/ highly qualified employee/ officer
	num_dependents	Continuous	1∼2
	sex	Binary	Female, Male.
	savings_status	Categorical	<100 DM, 100 <= ... <500 DM, 500 <= ... <1000 DM, >= 1000 DM, unknown/ no savings account
Label	Credit risk	Binary	good,bad

Bank			
	Name	**Type**	**Values**
Attributes	job	Categorical	admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown
	marital	Categorical	divorced, married, single, unknown
	education	Categorical	basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown
	default	Categorical	no, yes, unknown
	balance	Continuous	-8019∼102127
	housing	Categorical	no, yes, unknown
	loan	Categorical	no, yes, unknown
	campaign	Continuous	1∼63
	previous	Continuous	0∼275
	age	Continuous	-
	poutcome	Categorical	failure, nonexistent, success
Label	deposit	Binary	yes, no

Adult Census				
		Favor	Unfavor	Total Sample Size m
privileged	Orig	6669	14688	-
unprivileged	Orig	1158	9140	31655
	SMOTE	*4149*	9140	34646
	SM+ENN	*3521*	*7432*	32310
	SM+TL	*4107*	*9098*	34562
	SM+TL*	*4130*	*9098*	34585

German Credit				
		Favor	Unfavor	Total Sample Size m
privileged	Orig	418	150	-
unprivileged	Orig	79	53	700
	SMOTE	*148*	53	769
	SM+ENN	*139*	*12*	719
	SM+TL	*136*	*41*	745
	SM+TL*	*136*	*49*	753

COMPAS				
		Favor	Unfavor	Total Sample Size m
unprivileged	Orig	1042	1171	-
privileged	Orig	897	584	3694
	SMOTE	897	*1008*	4118
	SM+ENN	*561*	*586*	3360
	SM+TL	*896*	*1007*	4116
	SM+TL*	-	-	-

Bank				
		Favor	Unfavor	Total Sample Size m
privileged	Orig	3425	27302	-
unprivileged	Orig	213	707	31647
	SMOTE	213	*1704*	32644
	SM+ENN	*24*	*1351*	32102
	SM+TL	*148*	*1639*	32514
	SM+TL*	*205*	*1639*	32571

Table A.2: Datasets Tranining Sample Size