A EXTENDED RESULTS

A.1 Running Example

This example demonstrates the performance of DataSift during an experiment using a limited data pool ($|\mathcal{D}| = 50$) derived from the ACSIncome dataset [23]. The data pool was partitioned into an optimally determined number of clusters, with a mini-batch size of two instances. Table 1 details each step of the DataSift algorithm throughout the process, while Figure 10 presents the corresponding outcomes alongside comparisons to baseline methods. Despite the tiny data pool, DataSift effectively reduces existing unfairness by approximately 60%. Notably, DataSift consistently makes informed, fairness-aware selections during training, whereas baseline approaches exhibit limited ability to address the fairness issue.

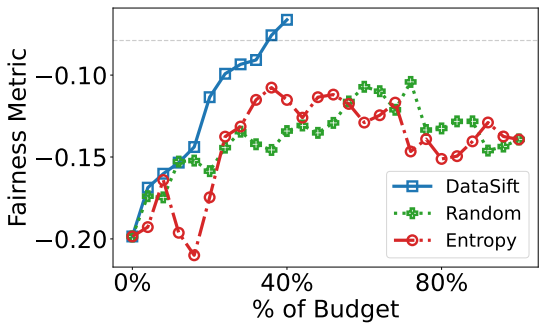


Figure 10: An example showcasing the performance of DataSift on a tiny sample of the ACSIncome dataset. $|\mathcal{D}| = 50$, $|K| = 2$.

A.2 Ablation Analysis

A.2.1 Effect of β. In Equation 1, the parameter β governs the trade-off between fairness and accuracy in shaping the reward function for each partition. Specifically, a lower value of β increases the weight assigned to fairness, encouraging the selection of instances that help reduce bias in the model. In contrast, a higher β emphasizes accuracy, favoring instances that contribute to performance improvement on the prediction task.

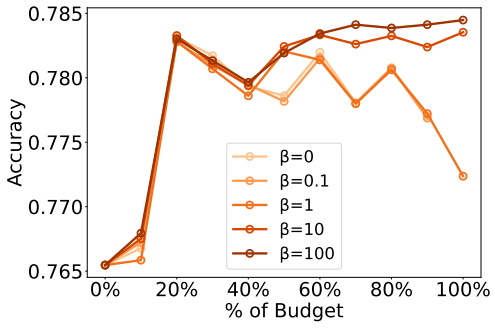


Figure 11: Effect of β on accuracy

Figure 7b illustrates that fairness steadily declines as β increases, confirming that the model becomes less fairness-aware when accuracy is prioritized. Figure 11 presents the corresponding accuracy trend, which improves with higher values of β. These results validate the intended behavior of the reward function: increasing β improves predictive accuracy but comes at the cost of slightly reduced fairness.

A.3 Different Fairness Metric

Algorithmic group fairness mandates that individuals belonging to different groups must be treated similarly. There are several fairness metric such as statistical parity (a.k.a. demographic parity), predictive parity, and equalized odds [16, 42, 59]. In our primary experiment, we consider statistical parity where DataSift outperforms the baselines. However, our proposed approach also performs well with other metrics. Figure 12 reproduces the same experiment as Figure 3 but for the fairness metric predictive parity. Predictive parity requires that a model's precision is equal across groups. Formally, for all protected groups $A = a, b$,

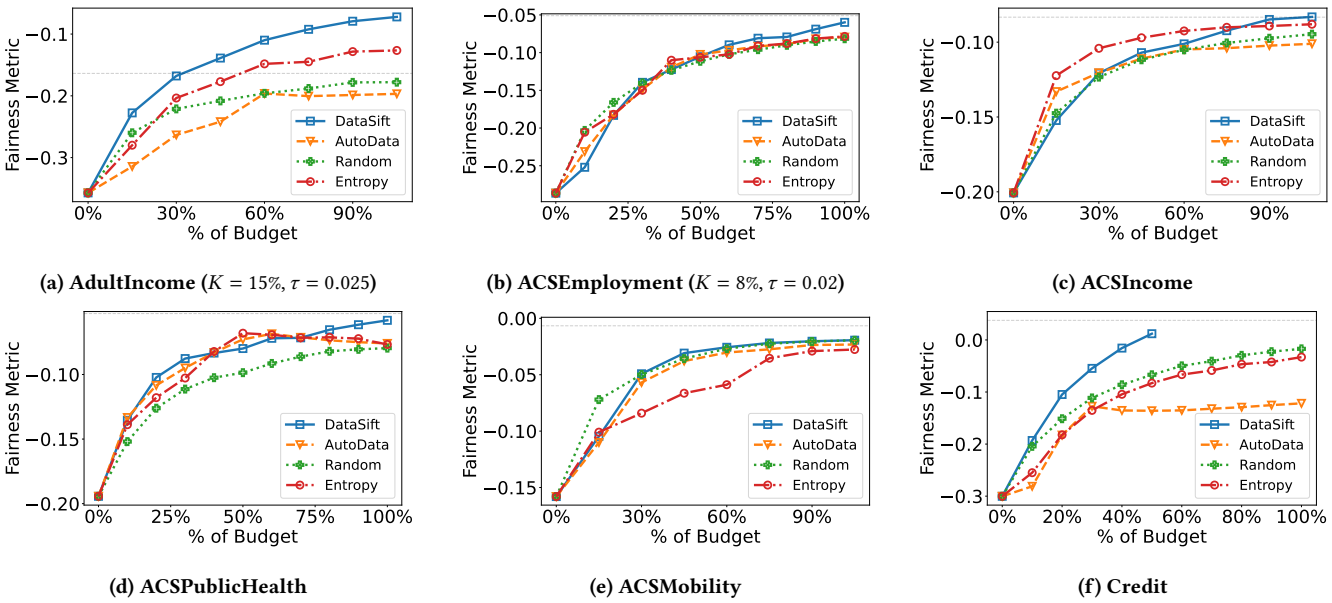
$$P(Y = 1 \mid \hat{Y} = 1, A = a) = P(Y = 1 \mid \hat{Y} = 1, A = b),$$

where Y is the true label and $\hat{Y}$ is the predicted label. This ensures that, among individuals predicted positively, the proportion correctly classified is the same across groups. In Figure 12, across all datasets, DataSift reaches fairness levels similar to or better than those achieved by using the full data pool (shown as the black dotted line in Figure 12), while using only a small portion of the data. In particular, DataSift is able to reduce 50% to 100% of the existing bias in the model by carefully selecting just 20% of the data pool, $|\mathcal{D}|$. In every case, it outperforms the baseline methods.

Now, we extend the results for DataSift-Inf in Figure 13, reproducing the same experimnt as Figure 5, but for the fairness metric predictive parity. As shown in Figure 13, DataSift-Inf consistently outperforms DataSift and Inf across all datasets. Notably, it achieves a fair model while utilizing no more than 60% of the allocated budget, outperforming all baseline methods in terms of both fairness and efficiency. In every case, DataSift-Inf reduces 100% of existing bias with significantly less data. Furthermore, it terminates the evaluation process early upon meeting the predefined fairness threshold, demonstrating its computational efficiency.

A.4 Accuracy

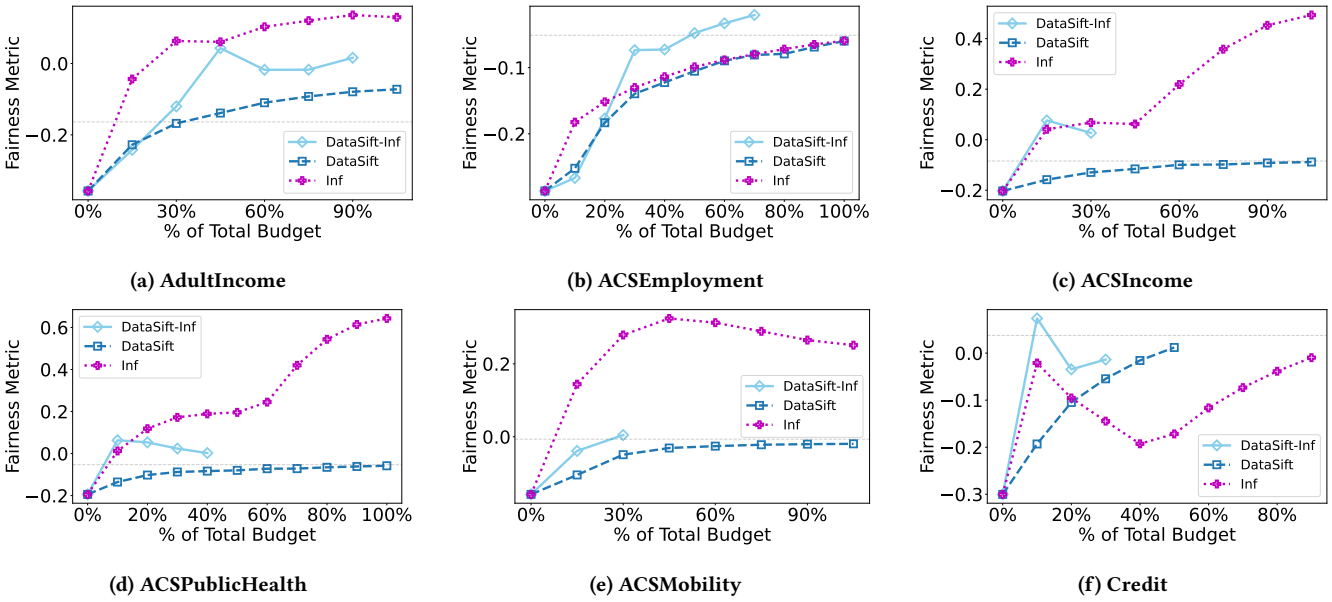
As shown in Table 3, using DataSift and DataSift-Inf, model accuracy either improves or preserves the initial accuracy for most datasets. For ACSEmployment, ACSPublicHealth, ACSMobility, and Credit datasets, DataSift is the top performer–improving the accuracy significantly. Slightly lower than DataSift, but DataSift-Inf improves the fairness in most of cases, otherwise preserves the accuracy. However, for DataSift-Inf we observe a slight drop in accuracy for the ACSIncome dataset. Although this decline is considerable with regard to fairness improvement, it can be addressed by increasing the budget size and allowing more data to be added. Note that the other methods, while preserving accuracy, do not ensure model fairness. Originally tailored to improve model accuracy, AutoData shows similar accuracy as the other methods because of the updated constraint that checks for improvement in fairness rather than accuracy. We also observe some extreme scenarios for Inf, which greedily exploits the most influential data points and fails to preserve fairness and accuracy in the process.



(a) **AdultIncome** ($K = 15\%, \tau = 0.025$) (b) **ACSEmployment** ($K = 8\%, \tau = 0.02$) (c) **ACSIncome**

(d) **ACSPublicHealth** (e) **ACSMobility** (f) **Credit**

Figure 12: Comparing DataSift-Inf, our MAB-approach based on data valuation, with DataSift and Inf to highlight the importance of data valuation in acquiring data that rapidly improves model fairness. DataSift-Inf exhibits the most improvement in fairness with the least amount of additional data added. The black dotted line indicates $\mathcal{F}_{\mathbf{D}_{train} \cup \mathcal{D}}$. Budget $|B| = .2 * |\mathcal{D}|$ and $|K| = .1 * |B|$.



(a) **AdultIncome** (b) **ACSEmployment** (c) **ACSIncome**

(d) **ACSPublicHealth** (e) **ACSMobility** (f) **Credit**

Figure 13: Comparing DataSift with baselines (AutoData, Random, and Entropy) to highlight the effectiveness in achieving fairness. DataSift consistently outperforms the other methods in improving fairness for most of the datasets. The black dotted line indicates ultimate fairness if the entire data pool is added (i.e., $\mathcal{F}_{\mathbf{D}_{train} \cup \mathcal{D}}$). Budget $|B| = .2 * |\mathcal{D}|$ and $|K| = .1 * |B|$.

Dataset	Classifier	Methods						
		Initial	Random	Entropy	AutoData	DataSift	Inf	DataSift-Inf
AdultIncome	Logistic Regression	0.73	**0.79**	0.78	0.73	0.76	0.75	0.77
	SVM	0.72	**0.79**	0.78	0.73	0.76	0.76	0.77
	Neural Network	0.75	**0.81**	0.79	0.79	0.76	0.73	0.77
ACSIncome	Logistic Regression	0.77	0.78	**0.78**	0.78	0.76	0.50	0.76
	SVM	0.77	0.78	**0.78**	0.78	0.74	0.59	0.77
	Neural Network	0.77	0.79	**0.79**	0.75	0.75	0.54	0.74
Employment	Logistic Regression	0.69	0.75	0.75	0.73	**0.76**	0.72	0.73
	SVM	0.69	0.74	0.75	**0.76**	**0.76**	0.65	0.73
	Neural Network	0.80	0.78	**0.80**	0.79	**0.80**	0.72	0.78
PublicHealth	Logistic Regression	0.65	**0.69**	0.68	0.68	0.68	0.47	0.68
	SVM	0.63	0.68	0.68	**0.69**	**0.69**	0.49	0.55
	Neural Network	0.69	0.70	**0.71**	0.67	0.65	0.48	0.68
Mobility	Logistic Regression	0.77	0.77	**0.77**	0.77	**0.77**	0.45	0.75
	SVM	0.77	0.76	**0.77**	0.77	**0.77**	0.75	0.76
	Neural Network	0.77	0.77	**0.77**	0.77	**0.77**	0.69	0.74
Credit	Logistic Regression	0.73	0.93	**0.93**	0.93	**0.93**	0.72	0.91
	SVM	0.73	0.93	**0.93**	0.93	**0.93**	0.88	0.84
	Neural Network	0.93	0.93	**0.93**	0.93	**0.93**	0.93	0.93

Table 3: Table presents the observed accuracy of a model trained after adding additional data from the data pool as determined by a given method. Bold values denote the highest accuracy, while underscored values signify the second-highest accuracy.