Evaluation of the quality of synthetic dataset for fraud detection

Ali Valiyev

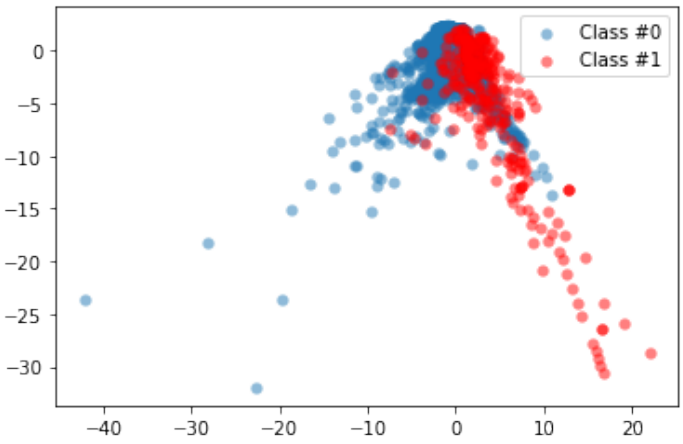
April 2022

1 Introduction

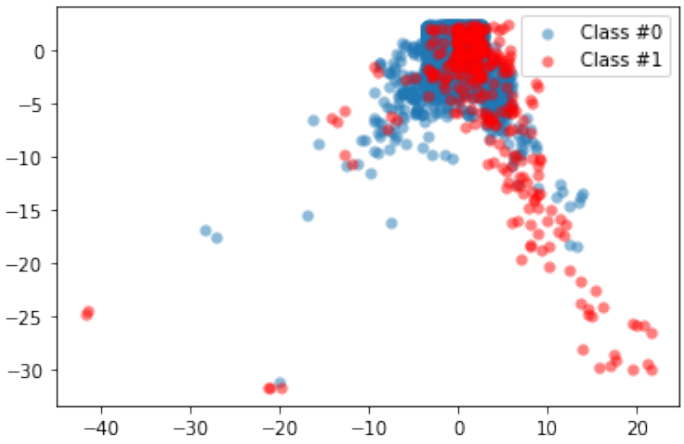
The objective of our experiment is to analyze the performance of Random Forest, Naïve Bayes, Logistic Regression, and K-nearest neighbors machine learning models for evaluation of the utility of synthesized data for fraud detection. We will use the accuracy, AUROC, precision, recall, and f1-scores to evaluate the model's performance.

Let's start the analysis with the exploration of the dataset. Using these scatter plots, we can visualize fraud and no fraud distributions of raw, SDV TVAE, and DS correlated attribute datasets. As can be noticed from these plots, the closest distribution to the raw dataset is synthetic data synthesized with DS with correlated attributes. However, the distribution of data points of the SDV TVAE synthetic dataset is distinctly different from the raw dataset.

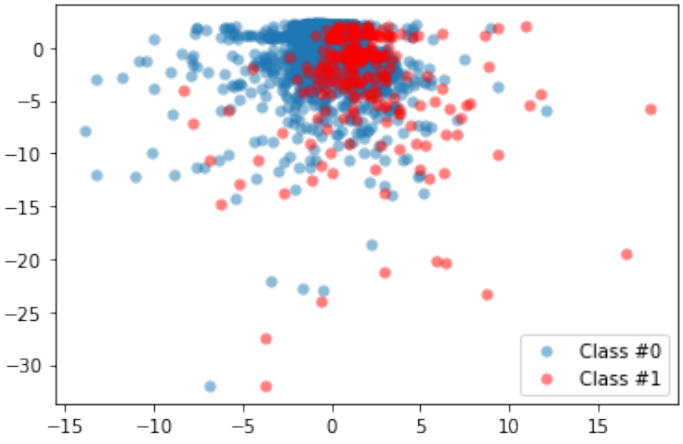
In the following chapters, we will continue our analysis with the evaluation of the performance metrics of different machine-learning prediction algorithms on raw and synthesized data.



(a) Scatter plot of a raw dataset



(b) Scatter plot of a synthetic dataset(DS correlated)



(c) Scatter plot of a synthetic dataset(SDV TVAE)

2 Fraud data characteristics

Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit or debit card. Usually, the intention is to obtain goods or services or make a payment to some account in a fraudulent fashion. The dataset used in this study contains

7300 credit card transaction data points; of each, 7000 are non-fraudulent, and 300 are fraudulent. As can be noticed from the number of observations in each category, the unique characteristics of fraud data are its imbalance property. Therefore, the goal of this study and hence, the next sections of the report is to investigate the applicability of synthetic data generators for fraud (imbalanced) data generation by comparing the performance of a range of standard machine learning algorithms on raw and generated synthetic datasets.

3 Definitions

In this section, we will define the performance metrics used in our experiment and explain their meanings in the context of fraud detection. Also, we will be using these words in the formulas: True Positive, True Negative, False Positive, and False Negative.
Meanings of True Positive, True Negative, False Positive, and False Negative in the context of fraud detection:
True Positive - Fraudulent transactions the model predicts as fraudulent.
True Negative - Normal transactions the model predicts as normal.
False Positive - Normal transactions the model predicts as fraudulent.
False Negative - Fraudulent transactions the model predicts as normal.
1) AUROC is the area under the receiver operating characteristic curve. AUROC is a performance metric for "discrimination": it tells us about the model's ability to discriminate between cases (fraud) and non-cases (non-fraud examples.)
2) Precision answers the question: out of all the transactions predicted to be fraudulent, what percentage were actually fraudulent?

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

3) Recall answers the question: out of the fraudulent transactions, what percentage of these are correctly identified by our model?

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

4) The F1 score combines Recall and Precision into one metric as a weighted average of the two. Unlike Recall and Precision individually, F1 takes both false positives and false negatives into consideration.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4 Accuracy

We can see from Figure 2 that the accuracy scores of all machine learning models are higher than 96%. More specifically, the Random Forest model for the raw dataset, synthetic dataset generated by DS correlated, and synthetic dataset generated by SDV TVAE algorithm have 99.2%, 99.0%, and 98.8% accuracies, respectively. As can be seen from the figure, the relatively worse result compared to other models has the Naive Bayes model. It has 96.9% accuracy for synthetic DS correlated and 96.0% for synthetic SDV TVAE. We need to point out that the performance of all these four models on the DS correlated synthetic dataset is higher than the synthetic dataset generated by SDV TVAE. However, since our dataset is imbalanced, the high accuracy scores can be misleading. Since, in our dataset, the number of non-fraud data points is much higher than the number of fraud elements. Therefore, in the following sections, we will look at other performance metrics to gain more insights from synthetic data.

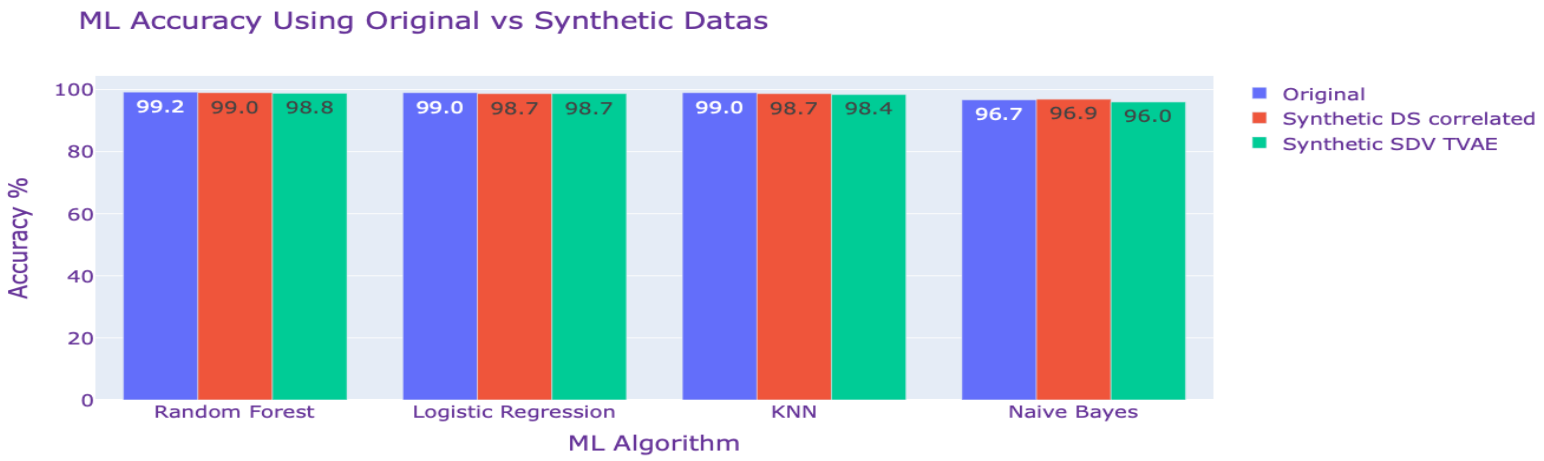


Figure 2: Accuracy score

5 Precision, Recall and F1-score

Figure 3 shows the Precision scores of models on original and synthetic datasets. For the DS correlated synthetic data Random Forest, Logistics Regression, and K-nearest neighbors can capture 93% of fraudulent transactions, but Naive Bayes has only a 63% precision rate. We also need to point out that Random Forest and K-nearest neighbors model has a 99% precision score which is exceptional. We can conclude from the precision scores that overall, all models on both synthetic data, especially on DS correlated, have very high precisions on capturing frauds, except the Naive Bayes model.

From Figure 4, we can observe that the recall score of Logistic Regression and K-nearest neighbors models on DS correlated have 79 percent accuracy, which means that (79%) of fraudulent transactions are identified by these models. Moreover, we can see that the values of the DS correlated, without differential privacy, are high for each of the five models, sometimes even higher than the scores of the model trained with the original data. In this context, we have to stress the fact that the dataset synthesized by the DS with $\epsilon = 0$ is much closer to the original than the dataset synthesized by the SDV TVAE.

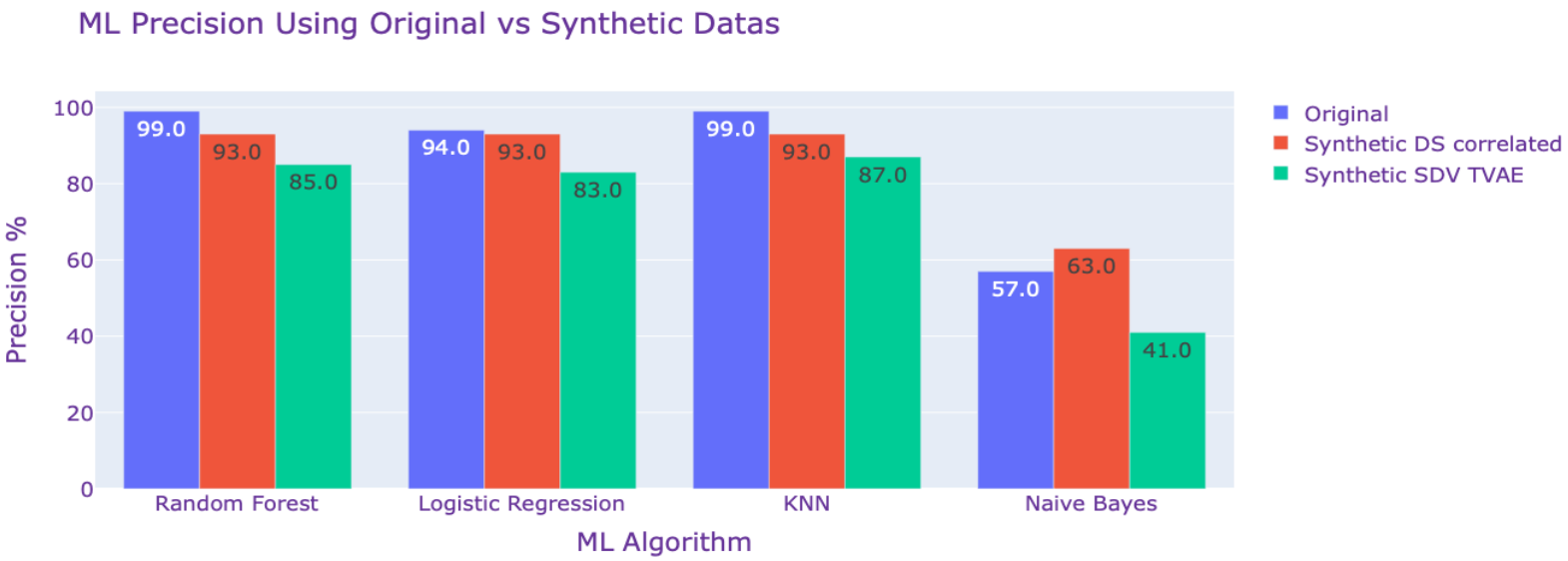


Figure 3: Precision

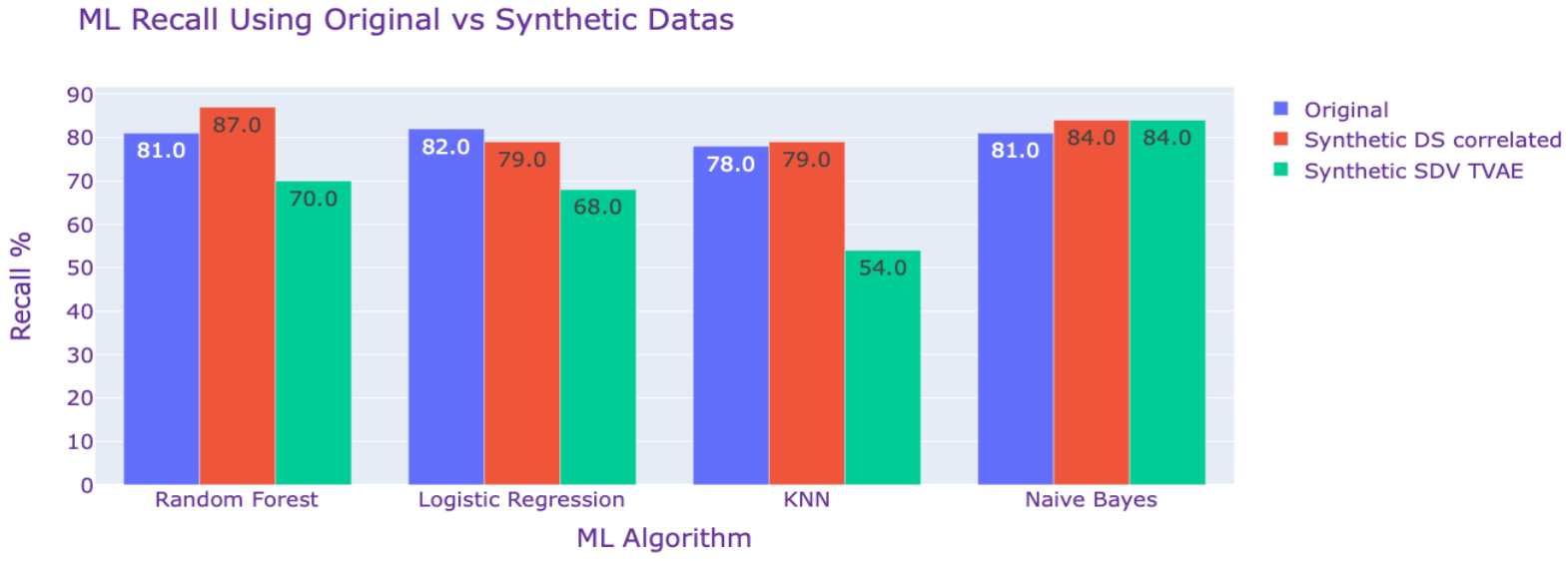


Figure 4: Recall

As we defined in section 3, F1-score takes false positives and false negatives into consideration, i.e., standard transactions that the model predicted as fraudulent and fraudulent transactions that the model predicted as normal. So we can say that F1 scores can give more insights into the synthetic data. From Figure 5, we observe that all models have high F1 scores on DS correlated synthetic data than on SDV TVAE synthetics data. Furthermore, Random Forest and Naive Bayes have (90%) , and (72%) performance on synthetic data generated by DS correlated, where these model's F1-score on the raw dataset is (89%) and (67%).

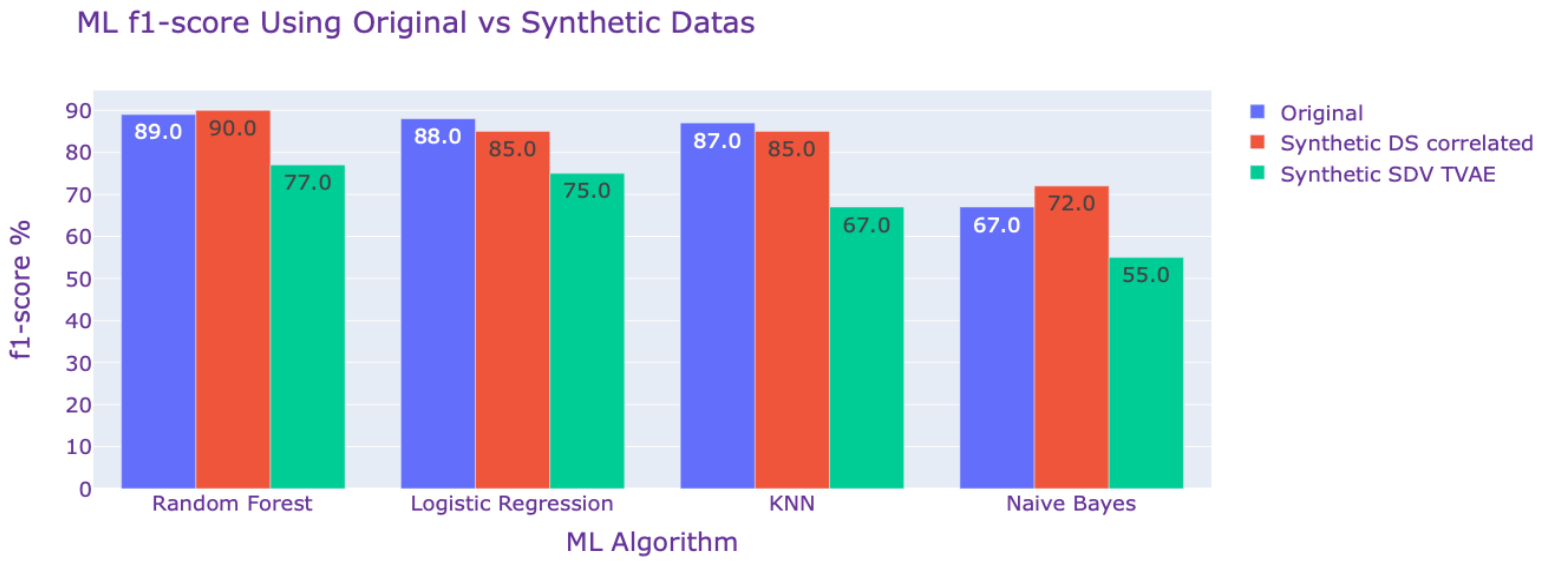


Figure 5: F1-score

6 AUROC

We continue discussing the area under the receiver operating characteristic performance metric(AUROC). The high AUROC means that the model's performance is better at distinguishing between fraud and no fraud classes. From Figure 6, we can observe that the scores of all models on synthetic SDV TVAE are higher than the percentages on synthetic DS correlated synthetic data. In addition, from the AUROC metric scores, we can see that all models have very similar performance for both raw and synthetic datasets. However, like in section 4 in our accuracy score analysis, the imbalance property of the dataset also affects to the AUROC score.

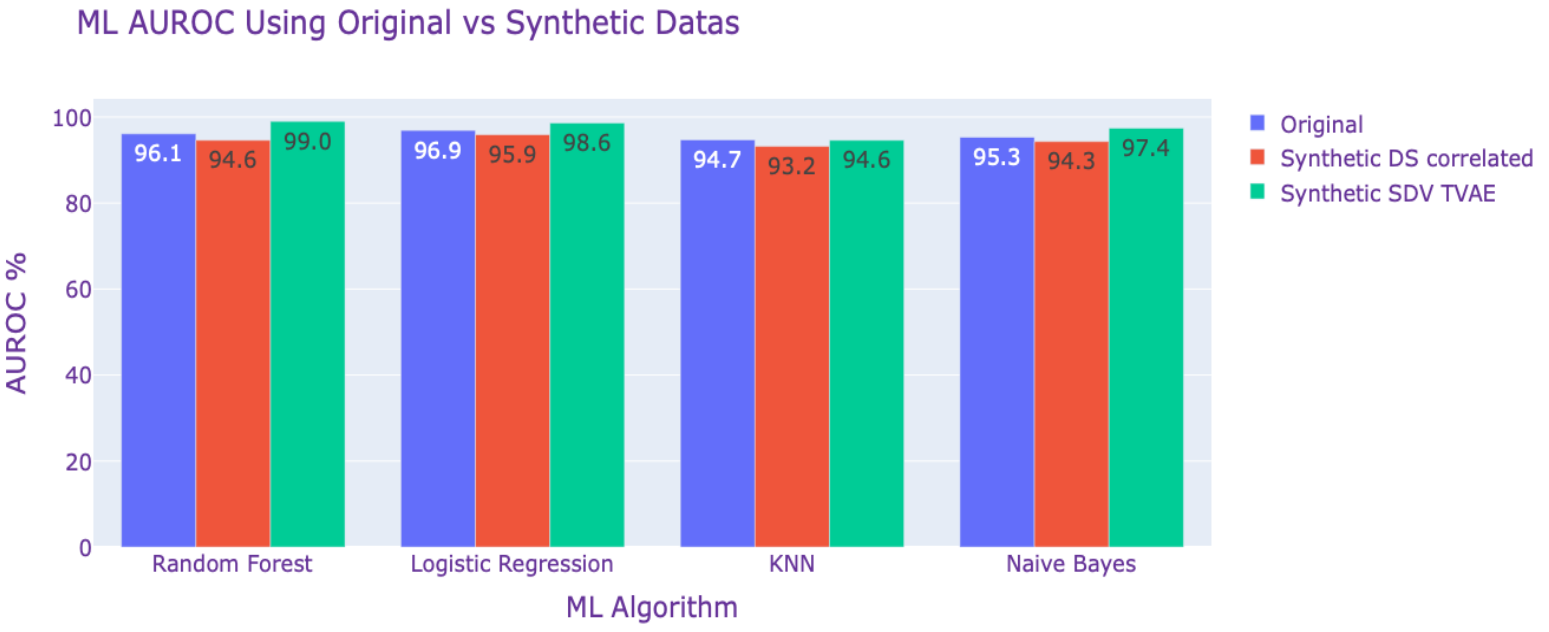


Figure 6: AUROC

7 Conclusions

Based on the analysis presented, first, we need to point out that for fraud detection, classical metrics such as accuracy and AUROC cannot capture the actual fraud identification rate due to skewness in instances for each class. Thus, we can conclude that metrics that are suitable for the detection of both classes of imbalanced data are precision, recall, and f1-scores. Furthermore, we can deduce that the data generated by DS correlated attribute with disabled differential privacy is more suitable for fraud detection in supervised machine learning settings. It is recommended to use the Logistic Regression, K-nearest neighbors, and especially Random Forest models for unbalanced data because their performance metrics such as precision, f1-score, and recall on DS correlated attribute synthetic datasets have higher accuracy scores and closer results to the raw datasets.