National College of Ireland

Generation of synthetic examples for imbalanced tabular data

MSc Research Project
Data Analytics

Nirav Bharat Gala
Student ID: 21125261

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nirav Bharat Gala
Student ID:	21125261
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	20/12/2022
Project Title:	Generation of synthetic examples for imbalanced tabular data
Word Count:	2176
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Generation of synthetic examples for imbalanced tabular data

Nirav Bharat Gala
21125261

Abstract

It is a laborious process to grant a loan because it requires extensive verification and confirmation. Banks and lenders must evaluate the credit risk involved with each loan application in order to prevent defaults. Data analytics and Machine learning techniques could be used on historical data to predict loan defaults and enable loan officers to make informed decisions. Data being used in this processes is often class imbalanced where defaulters are the minority classes. Many techniques have been proposed to balance out the classes, but it is not known what technique works best for credit defaults. This report presents a detailed comparison of the most important techniques for class imbalance to develop a robust binary classifier. Generative adversarial networks (GAN) and Synthetic Minority Oversampling (SMOTE) are the two most important techniques for the generation of synthetic data. A detailed comparison between GAN and SMOTE is presented in this report. The recall metric was employed to evaluate the techniques because it represents the model's ability to identify potential defaults. Although both techniques compared relatively well on the generation of synthetic data for loan default, we will show that SMOTE outperforms GAN in terms of recall.

1 Introduction

Financial entities like banks and lenders get revenue by charging interest and fees on approved loans. Lending involves a lot of risk, including the possibility that the borrower will not be able to pay back the loan. The term "default" refers to debtors who are unable to repay loans, and it results in a financial loss for the lenders. It is crucial to determine the level of risk that each applicant has in order for the lender to decide whether or not to approve that applicant's loan application in order to avoid this situation.

This financial entities want to predict who might default on a consumer lending product. They have information about previous client behaviour based on what they have observed. So when they get new customers, they want to know who is riskier and who is not. This process is often called as predictive analytics where historical data is leveraged to build machine learning models. These machine learning models are essential binary classifiers that predict whether the loan would default or not. The data on which these models are fitted is often class imbalanced where the number of observations in one class is significantly more than the other. Defaulters constitute the minority class and the applicants who repaid their loans are the majority class. The minority class, by definition, has few examples, making prediction more difficult. This indicates that

learning the traits of examples from this class and differentiating them from instances from the majority class will be more difficult for a model (or classes).

The minority class may get overwhelmed by the amount of examples from the majority class (or classes). The study by (Longadge and Dongre; 2013) states that the majority of machine learning methods for classification predictive models are developed and tested on issues where it is assumed that the classes are distributed equally. Because of this, a model's naive application may only concentrate on learning the traits of a large number of observations, ignoring the examples from the minority class, which is actually more interesting and whose predictions are more important.

The research question

The proposed research tackles this class imbalance problem by generating synthetic data. There has been extensive research in deep learning to generate synthetic data using generative adversarial networks (GAN). GANs can be used to generate images as well as tabular data. Generative adversarial networks and the Synthetic Minority Oversampling Technique (SMOTE) are probably the two most popular techniques for synthetic data generation and their impact is assessed based on certain metrics. This brings us to our research question:

> *How do GAN and SMOTE techniques compare in the generation of synthetic tabular data for the prediction of loan defaults?*

While there are classification techniques for imbalance data, the primary research objective of this report is to address the class imbalance itself. We will show ways to generate synthetic data for default loans and then apply a classified on the balanced data. GAN and SMOTE are the two techniques that would be used to generate synthetic data and detailed comparison between them will be performed to determine their effectiveness. In addition to that, this study implements the entire machine learning workflow that involves exploratory data analysis, data pre-processing, model building and evaluation.

Structure of document

The remainder of the report is organised as follows. Section 2 presents the relevant research studies and reviews the methodologies in it. Section 3 describes how the CRISP-DM methodology is adopted in the proposed research. In Section 4.1, the workflow followed in this research is described. The two techniques namely generative adversarial networks and SMOTE are presented in Section 5, while experiments performed and results obtained appear in Section 6. Finally, we provide conclusions and discussions of future research directions in Section 7.

2 Related Work

Class imbalance is found in the datasets due to intrinsic property of the problem or due to limitations to obtain data such as cost, privacy and large effort. The study by (Abd Elrahman and Abraham; 2013) elaborates that this problem is an hindrance to machine learning algorithms and their performance in making predictions. Sampling methods such as under-sampling the majority class and oversampling the minority class

are suggested to solve the problem however the author fails to elaborate how under-sampling can lead to loss of useful information and patterns in data whereas oversampling can lead to over-fitting since it replicates the existing samples. Furthermore,(Ramyachitra and Manikandan; 2014) the evaluation metric accuracy would be biased to the majority class and so is not a reliable measure to evaluate models. Metrics, such as precision, recall and F1 scores, are more meaningful when reporting imbalanced classification problems.

The performance of several re-sampling algorithms to cope with imbalanced data sets is examined in the study by (García et al.; 2012), along with the effects of the imbalance ratio and the classifier. The goal of the study is to assess how learning is impacted when various re-sampling algorithms convert the initially unbalanced data into fictitiously balanced class distributions. Over-sampling the minority class consistently outperforms under-sampling the majority class when data sets are strongly imbalanced, whereas there are no significant differences for databases with a low imbalance, according to experiments over 17 real data sets using eight different classifiers, four re-sampling algorithms, and four performance evaluation measures. For oversampling the minority class, synthetic data needs to be generated.

Generative Adversarial Networks (GAN)

The review presented by (Gonog and Zhou; 2019) elucidates that the structural inspiration for GANs comes from two-person zero-sum games in game theory (i.e. the sum of the two people interests is zero, and the gain of one side is exactly what the other side loses). For each player in the game, one generator and one discriminator are set up. The generator's goal is to generate new data samples while learning as much as it can about the prospective distribution in the real data samples. The discriminator is a binary classifier whose goal is to ascertain if the input data comes from the generator or the actual data. The two players must continually enhance their capacity to generate and distinguish in order to win the game.

The probability distribution of a dataset is implicitly learned by generative adversarial networks (GANs), which can then extract samples from the distribution as stated by (Xu and Veeramachaneni; 2018) in this research. Tabular GAN (TGAN), a generative adversarial network that can produce tabular data like medical or academic records, is described in this report. With the help of deep neural networks, TGAN produces both discrete and continuous variables as well as high-quality, fully synthetic tables. TGAN is compared with other synthetic data generation techniques like Gaussian Copula and Bayes Network on the adult census, KDD99 and covertype dataset. Model is fitted on synthetic data and tested on test data which was set aside in the beginning. The results these models achieved are summarized in the following figures. As opposed to 24.9% for GC and 43.3% for BN-Co, the average performance gap between real data and synthetic data for TGAN is 5.7%.

Method	Real	GC	BN-Id	BN-Co	TGAN
DT					
max_depth = 10	74.65	48.61	32.26	32.24	68.70
max_depth = 20	75.11	48.64	31.16	31.77	64.42
SVM	71.30	-	-	25.69	67.77
RF					
max_depth = 10, estimators = 10	59.04	-	-	-	51.42
max_depth = 20, estimators = 10	70.95	-	-	32.26	65.89
AdaBoost	74.10	-	-	32.27	70.08
MLP					
layer_sizes = (100,)	75.47	53.15	25.5	26.34	71.81
layer_sizes = (200, 200)	73.94	-	-	32.14	68.75

Figure 1: Comparison of Synthetic data generation techniques.Source:(Xu and Veeramachaneni; 2018)

	KDD99			covertype			
Model	Real	GC	TGAN	Real	GC	BN-Co	TGAN
DT							
max_depth = 10	97.75	58.34	90.14	77.43	46.10	48.76	69.27
max_depth = 30	97.35	56.46	80.58	90.82	36.83	46.21	58.88
SVM	93.64	56.15	94.56	70.97	46.30	48.76	67.94
RF							
max_depth = 10, estimators = 10	97.79	60.61	93.36	74.58	45.30	48.91	66.60
max_depth = 20, estimators = 10	97.81	56.46	92.33	85.13	46.78	48.85	69.33
AdaBoost	19.93	75.94	40.43	49.81	40.95	48.88	66.11
MLP							
layer_sizes = (100,)	97.48	56.15	95.91	60.32	47.82	48.86	61.24
layer_sizes = (200, 200)	96.08	56.14	66.38	84.11	46.97	48.79	68.80

Figure 2: Comparison on KDD99 and covertype dataset. Source: (Xu and Veeramachaneni; 2018)

The study developed by (Xu et al.; 2019) proposes CTGAN as a GAN-based technique for modeling the distribution of tabular data and selecting sample rows from the distribution. To overcome the non-Gaussian and multi-modal distribution in CTGAN, mode-specific normalization is developed. To address the imbalanced discrete columns, conditional generator and training-by-sampling is built. The study by (Figueira and Vaz; 2022) proposes statistical techniques to evaluate generated synthetic data and compare mean ,median and standard deviation of variables with real data.

The paper by (Ashrapov; 2020) implements CTGAN and proposes a very effective strategy to determine the quality of synthetic data. The model is fitted on imbalanced training data and compared with the model fitted on synthetic data. This exercise is performed on 9 datasets and there is not significant difference in the performance of these models for seven datasets where ROC-AUC score is used as the evaluation metric. This datasets belong to different domains and having different number of observations. Moreover each dataset has a mix of categorical and numerical features and the objective of each one of them is binary classification. So this methodology and technique to generate synthetic data was chosen in the research.

Synthetic Minority Oversampling (SMOTE)

In this study, (Chawla et al.; 2002) defines SMOTE as a approach which is a blend of undersampling the majority class and oversampling the minority class. Instead of oversampling with replacement,synthetic samples of minority class are created. SMOTE was performed on 9 datasets with varying degree of imbalance and compared with plain undersampling. SMOTE offers more relevant minority class examples for learning, enabling learners to create bigger decision areas and increasing the minority class's coverage and so machine learning algorithms after applying SMOTE deliver better performance for minority class in terms of precision and recall.

The study by (Alam et al.; 2020) compares the models built on imbalanced datasets with the models built on balanced ones. Results for imbalanced datasets indicate that the accuracy for the credit data for Taiwanese clients is 66.9%, the accuracy for South German clients is 70.7%, and the accuracy for Belgian clients is 65%. Using the other hand, on the credit datasets for Taiwan clients, South German clients, and Belgian clients, the results employing our proposed approaches greatly increase accuracy by 89%, 84.6%, and 87.1%, respectively. The results demonstrate that classifier performance on the balanced dataset is superior to that on the imbalanced dataset. Additionally, it has been found that data oversampling approaches perform better than undersampling ones.

A Loan default prediction model is developed by (Zhu et al.; 2019) on peer-to-peer lending dataset that was inherently class imbalanced. Machine learning algorithms like Logistic Regression, Support Vector Machine and Random Forest are applied after data preprocessing where random forest outperforms the rest. SMOTE is used to tackle the class imbalance problem and the random forest classifier gives recall and f1 score of 0.99 and 0.98 for the default class respectively.

Research Gap and Analysis

Although many research studies have applied generative adversarial networks techniques and SMOTE for synthetic data generation, there is a absence of comparative study between them. SMOTE has been compared with oversampling and under-sampling and has always managed to outperform these techniques whereas to the best of my knowledge, its performance with tabular generative adversarial networks (TGAN) is not compared. Moreover in the context of loan default prediction, the performance of TGAN in tackling class imbalance and building a robust binary classifier is not explored.

3 Methodology

This study will follow the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for development. It consist of six sequential stages as shown in the figure 3. Each stage is explained in the subsections below.

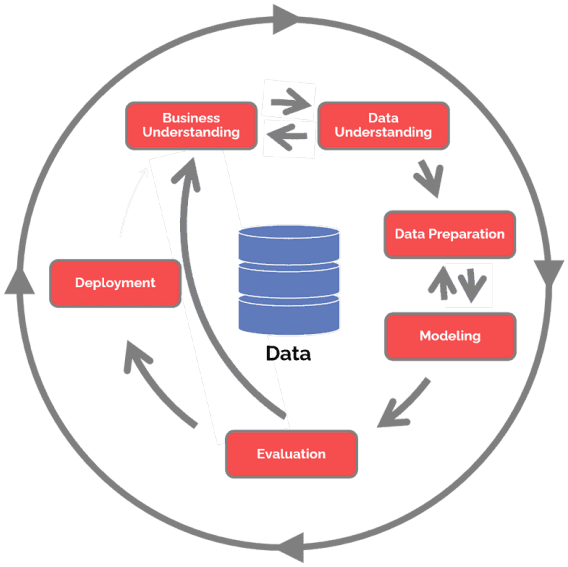

Figure 3: CRISP-DM methodology. Source `http://www.datascience-pm.com`

3.1 Business Understanding

This phase focuses on understanding the objectives and requirements of the project. Banks and lenders would greatly benefit from the proposed model since it would speed up the entire process of granting the loan from application to closing and help them in decision making. Data Mining and Machine learning techniques shall be used to determine the eligible loan applicants and the applicants most likely to default. This would further enable them to minimize losses.

3.2 Data Understanding

The dataset used in this research is sourced from an online open source data repository Kaggle and its called Bank Loan status dataset. It consists of 1,00,000 rows and 19 features where Loan_Status is the target feature and the rest are predictor features. The predictor features includes 4 categorical and 14 continuous features. Exploratory data analysis is performed that involves Uni-variate and Bi-variate analysis to determine relationship between variables and visualizations are created to discover insights in the dataset. The following Figure 4 shows that distribution of target variable and it is evident that the dataset is massively imbalanced. Loans that are paid are majority class and charged-off or default loans are the minority class.

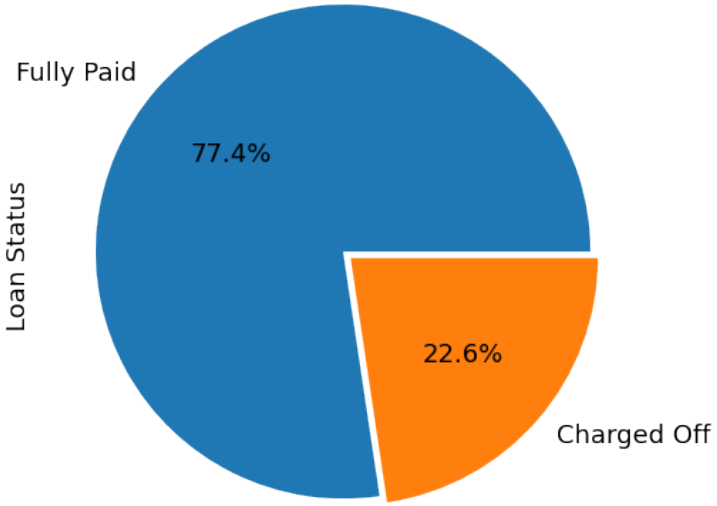

Figure 4: Class Imbalance in the dataset

3.3 Data Preparation

This phase involves data cleaning and data transformation. This dataset contained 10215 duplicate rows which were dropped. There were missing values for various features as shown in the table 1.

Table 1: Missing values in the dataset

Feature name	Missing Values	% of Total Values
Months since last delinquent	48337	53.8
Credit Score	19154	21.3
Annual Income	19154	21.3
Years in Current Job	3802	4.2
Bankruptcies	190	0.2
Tax Liens	9	0.0004
Maximum Open Credit	2	0.0001

There were different approaches taken to deal with each attribute. Months since last delinquent feature was dropped since it has more than 50% of missing values. Rows having missing values for bankruptcies, Tax Liens and Maximum Open Credit were dropped since the missing value percentage was very less and there would not be significant loss of information. For the rest of the variables,imputation is performed. Years in current job is a categorical variables and therefore the missing values are imputed by the most frequent value which is mode. For continuous variables Credit score and annual income,missing values are imputed by median. Another issue which was discovered that the credit score feature has values more than 850 which was practically impossible. This error was due to an extra zero at the right in the credit score value making it a 4 digit number and this was removed.

After dealing with missing values and data inconsistencies,label encoding is performed to transform categorical variables. Label encoding was preferred over one-hot encoding to escape the dummy variable trap and avoid multicollinearity. Features Loan-id and customer-id were also dropped since they were not adding any significant information in predicting the target variable.

3.4 Modelling

Prior to applying the machine learning algorithm, cross-validation will be performed to divide the dataset into training and test set. Classification algorithms such as Logistic Regression, Decision Trees and Random Forest shall be applied to predict if the loan application is default or not.

3.5 Evaluation

It is essential to compare models using performance metrics before deciding which model would help achieving the objective. Accuracy, Precision, Recall and ROC-AUC score are the extensively used metrics to evaluate performance of classification models, however accuracy would not be so significant in the context of this problem as stated by (Hossin and Sulaiman; 2015) because data is imbalanced and models would favor the majority class. Recall, Precision and ROC-AUC score of a model for the default class are more important since the objective is to identify the loans that are most likely to default.

4 Design Specification

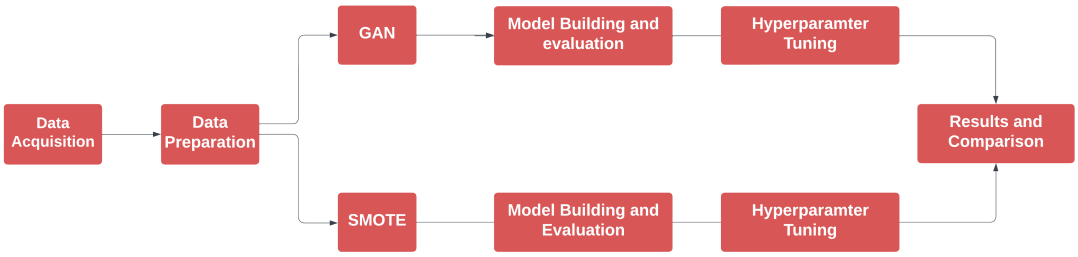

Figure 5: Workflow Diagram

After the data preparation stage explained in detail in the methodology section, the problem of class imbalance is resolved by generating synthetic data. First data is generated using Generative adversarial networks and the synthetic samples from the default class(minority) are added to the training data and reduce the class imbalance. Classification algorithms are applied to this data and evaluated based on metrics like Precision, Recall and ROC-AUC score. In the same way, SMOTE is used to solve class imbalance and model building is performed. In the end, performance of both the synthetic data generation technique is compared based on the performance of the models build after applying them.

4.1 GAN Architecture

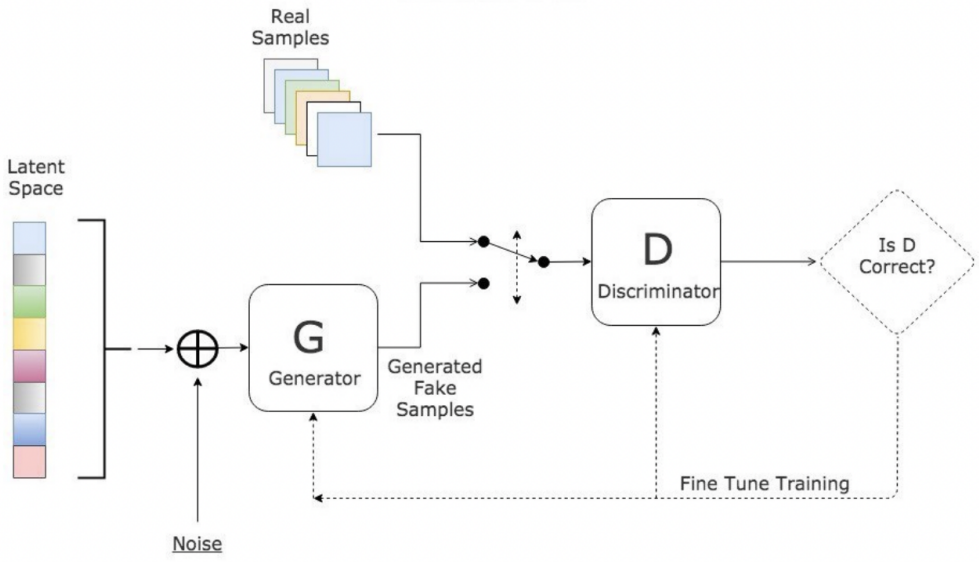

Figure 6: GAN architecture Source:(Ashrapov; 2020)

This network consists of two components which are essentially neural networks explained as under:-

- The generator gains the ability to produce credible data. The created samples serve as the discriminator's negative training examples.
- The discriminator gains the ability to discern real data from bogus data generated by the generator. The generator is punished by the discriminator when it generates improbable outcomes.

As training progresses, the generator generates data that is manifestly false, and the discriminator quickly learns to recognize it as such.

5 Implementation

The entire implementation of this research project is carried out using Python programming language because it offers very powerful and flexible libraries for data analysis, data visualization, synthetic data generation and machine learning. The tool used for implementation is Google Collab which provides a web-based environment to create and execute notebooks.

5.1 Data Exploration and Preparation

As discussed in the methodology section, uni-variate and bi-variate analysis is undertaken on the dataset using a library called Pandas profiler. This library enables to create a entire summary report of the dataset providing descriptive statistics for continuous variables and distribution of values for categorical variables.

Data cleaning is performed using Pandas where missing values are handled which is explained in detail in the methodology section. Data inconsistencies in certain attributes

is also dealt with and data visualizations like pie chart and bar charts are created using libraries like Matplotlib and Seaborn. After data is cleaned, label encoding is performed to handle categorical variables.

5.2 Synthetic Data Generation

The pre-processed data is still unbalanced where the defaulters are the minority class. To balance the dataset, synthetic data is generated using a library called tabgan which implements the tabular generative adversarial network (TGAN). To evaluate the quality of this synthetic data, models were fitted on this synthetic data and tested on the test data and the performance was not significantly lower than the real data. The synthetic records of the default class are added to the real data to increase the records of minority class and eventually make the distribution of the target classes even. Similarly, SMOTE is used to generate the synthetic data and balance the dataset. The implementation of SMOTE is performed using a library called imblearn and TGAN is implemented using library called tabgan. This balanced data has to be normalized to improve performance as stated by (Raju et al.; 2020) This balanced dataset is scaled using robustScaler since it minimizes the effect of outliers.

The below figures 7 and 8 display the distribution of values of variables in real data and synthetic data generated by GAN respectively. It is evident that synthetic data does not drift away from the real data for every categorical and numerical variable.

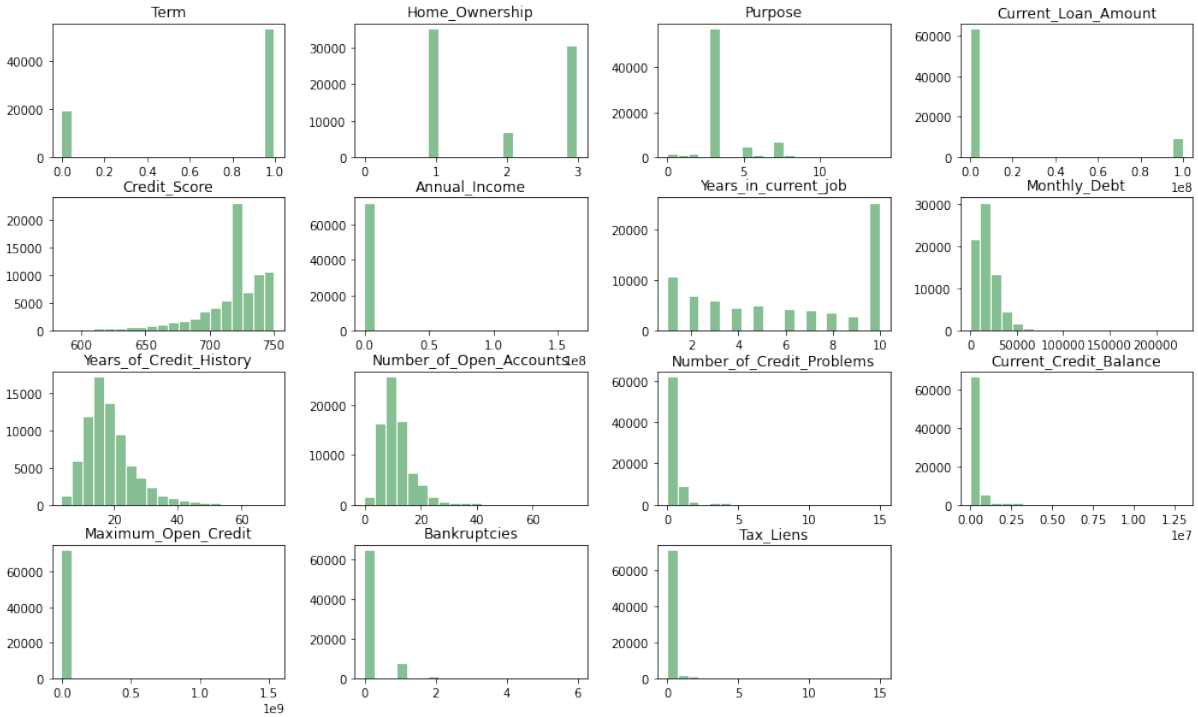

Figure 7: Real data

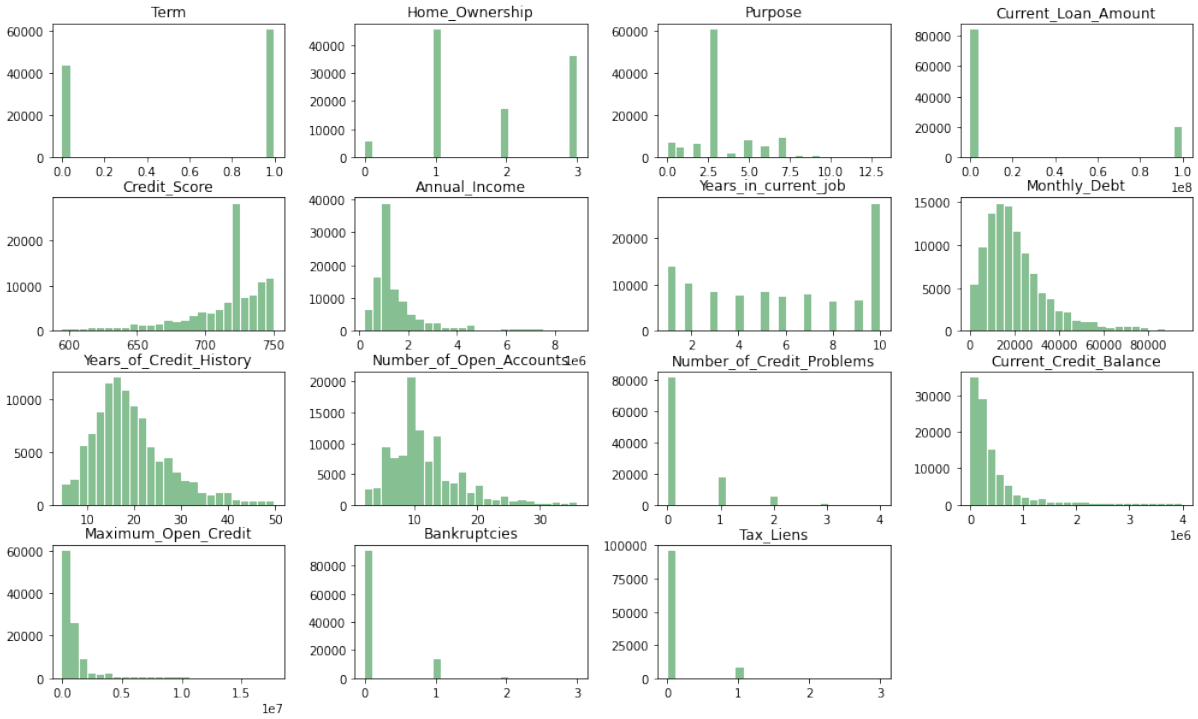

Figure 8: Synthetic Data generated by GAN

The X and Y axis are different for some variables because TGAN is unable to generate maximum values for certain variables. For example the range of values for variable Bankruptcies in real data is [0,7] while for generated synthetic data it is [0,3]. The synthetic data was evaluated based on two methods which are descriptive statistics and machine learning efficacy. The variation between the real and synthetic data was computed in terms of percentage change in descriptive statistics. The following table 2,3 and 4 displays the summary statistics for real data, synthetic data and the percentage difference for independent variables that were significant in predicting the target variable Loan Status.

Table 2: Descriptive statistics of real data

Statistic	**Credit Score**	**Annual Income**	**Years of Credit History**
Mean	719.93	1331976.22	18.24
Median	725	1169773	17
Standard Deviation	25.09	983125.71	7.03

Table 3: Descriptive statistics of Synthetic data

Statistic	**Credit Score**	**Annual Income**	**Years of Credit History**
Mean	715.5	1596454.8	19.19
Median	725	1169773	17.84
Standard Deviation	30.23	1227545.75	7.80

Table 4: Percentage change in values

Statistic	Credit Score	Annual Income	Years of Credit History
Mean	0	0	4.95
Median	0.6	19.85	5.21
Standard Deviation	20.48	24.86	11

It is evident from the above table that percentage change between synthetic data and real data is not very much. The change in median, mean and standard deviation is less than 5, 20 and 25 percent respectively. To ensure the underlying pattern in data is captured, machine learning efficacy is used. Models are fitted on real data and synthetic data and evaluated on the test set. The results obtained are displayed in the following table 5. The difference in accuracy obtained on test set between real and synthetic data should not be more than 5 percent.

Table 5: Machine learning efficacy

Machine Learning model	Real data	Synthetic data
Logistic regression	75	72
Decision Tree classifier	73	72
Random Forest Classifier	78	74

5.3 Modelling and Hyperparameter Tuning

Classification machine learning algorithms such as Logistic Regression, Decision Tree and Random Forest are applied on the scaled training data to build classification models. The performance of this models is evaluated based on their performance on test data. To optimize the performance of this models, hyperparamter tuning is performed using GridsearchCV. Since our objective is to prevent defaults, the metric of interest is Recall for the default class rather than overall accuracy of the classifier. The library used for model building and hyperparamter tuning is Sci-kit learn.

6 Evaluation

Based on the objective of this research, classification model is built on balanced and imbalanced data and comprehensive analysis of the classification report is performed eventually. There are three experiments carried out where one is on imbalanced data and other two are on balanced data. Synthetic data generation technique for balancing the dataset used in second experiment is GAN and in the third experiment, SMOTE is used to balance the data.

6.1 Experiment 1 : Modelling on Imbalanced Data

The first experiment involves model building and predicting on the imbalanced data. Positive class refers to the default or charged-off loans and negative class refers to the loans which are fully paid. Datasets were divided into training and test set consisting of 80 % and 20% data respectively. Models are fitted on the training set and evaluated on the

test set. Although, Logistic regression and RandomForest classifier deliver the accuracy of 75%, recall value is extremely less for this models. Also, Decision Tree classifier has the highest recall (35 %) for the default class. Recall is the most important metric in our research project since it denotes the ability of the model to identify defaults. Figure 9 visualizes the receiver operating characteristic of the three models built and displays the area under the curve.

Table 6: Classification metrics of models

Algorithm	Accuracy	Precision(default class)	Recall(default class)
Logistic Regression	75	49	4
Decision Tree Classifier	66	33	35
Random Forest Classifier	75	49	11

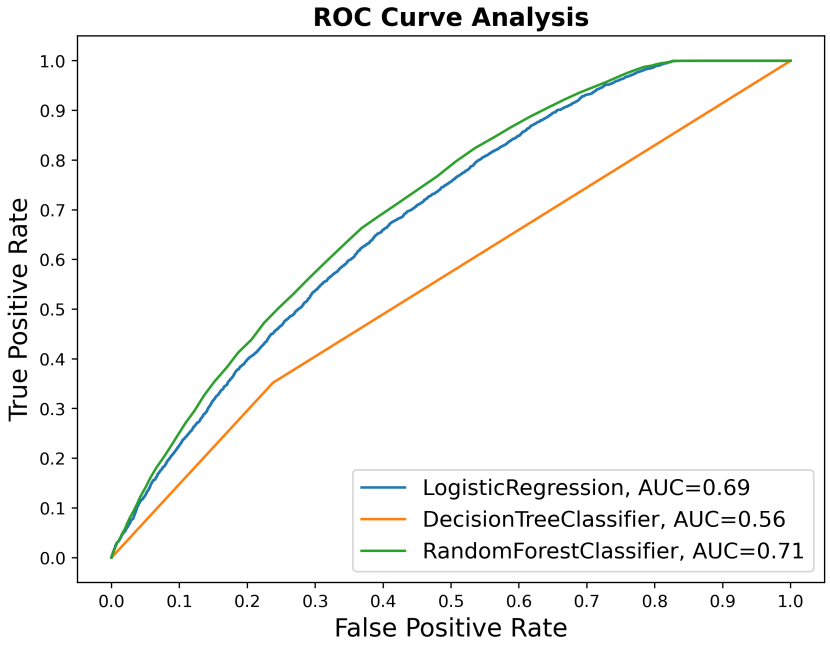

Figure 9: ROC-AUC Curve

6.2 Experiment 2 : Modelling on data balanced using GAN

The second experiment involves model building and predicting on the balanced data. Synthetic data is generated using GAN and the size of minority class which refers to defaults is increased by adding synthetic records. Models are fitted on the training set and evaluated on the test set. Logistic Regression has the least accuracy(59 %) and the maximum recall(62 %). In contrast to this, random forest has accuracy of 74% and delivers recall of merely 24%. Figure 10 visualizes the receiver operating characteristic of the three models built and displays the area under the curve

Table 7: Classification metrics of models

Algorithm	Accuracy	Precision(default class)	Recall(default class)
Logistic Regression	59	33	62
Decision Tree Classifier	66	33	43
Random Forest Classifier	74	44	24

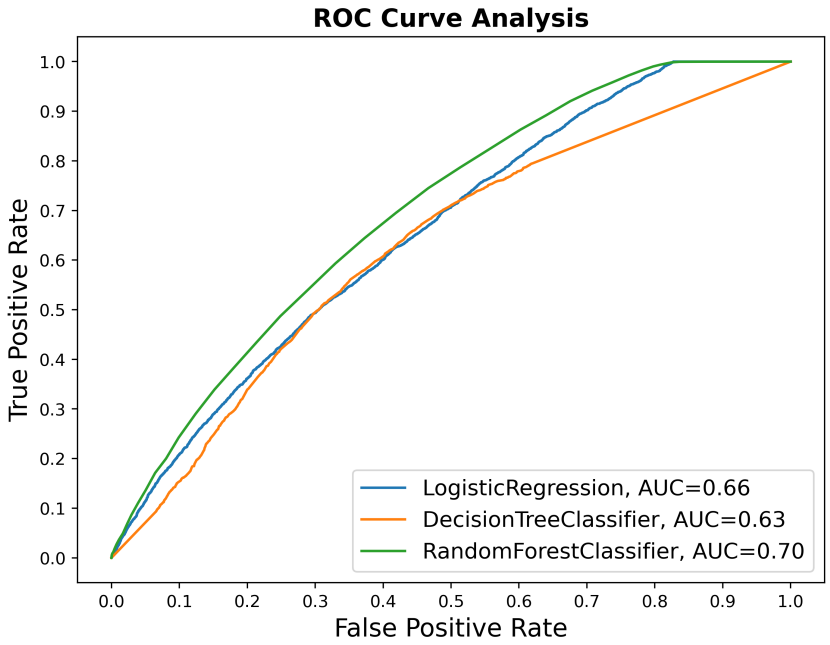

Figure 10: ROC-AUC Curve

6.3 Experiment 3 Modelling on data balanced using SMOTE

The third experiment involves model building and predicting on the balanced data. Instead of GAN, the data is balanced by using SMOTE. Models are fitted on the training set and evaluated on the test set. Logistic Regression has the least accuracy(56 %) and the maximum recall(74 %). In contrast to this, Decision Tree has the maximum accuracy of 63% and delivers the least recall among the three of 49%. Figure 11 visualizes the receiver operating characteristic of the three models built and displays the area under the curve

Table 8: Classification metrics of models

Algorithm	Accuracy	Precision(default class)	Recall(default class)
Logistic Regression	56	33	74
Decision Tree Classifier	63	34	49
Random Forest Classifier	58	33	65

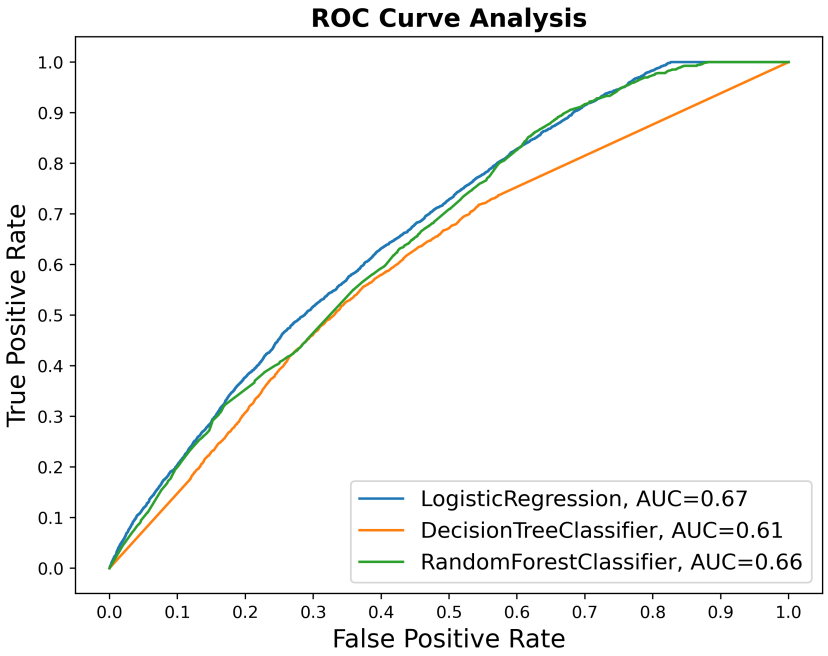

Figure 11: ROC-AUC Curve

6.4 Discussion

The results of first experiment which involves modelling on imbalanced data convey that models deliver good accuracy but poor recall for the default class. This is because the default class is the minority class and the model is overwhelmed by the records of majority class due to which its ability to detect potential defaults is very low. In the second experiment, the data is balanced using GAN, models are built and hyper-parameter tuning is performed where models are optimized for recall. This results in significant increase of recall for the default class for all the three algorithms as shown in Table 10 and minor decrease in overall accuracy of the models. When the data is balanced using SMOTE in the third experiment, there is significant increase in recall but significant decrease in accuracy as well for all the three algorithms. The results of the experiments imply that decision tree, random forest and logistic regression are sensitive to class imbalance. Moreover, it was observed that there is not much variance in ROC-AUC score as shown in Table 11 for all the three models in all the three experiments. When assessing a model's performance on a dataset with class imbalance, the ROC-AUC score is not always a reliable metric. This is due to the fact that the ROC-AUC score may be biased in favor of the majority class because it does not account for the relative sizes of the various classes in the dataset. In addition to this, the precision of the models reduces when the data is balanced because when hyperparameter tuning is performed , models are optimized for recall since the most important metric in the context of this problem statement is recall.

The limitation of this research study is that the quality of synthetic data generated by GAN is only evaluated by descriptive statistics and visually assessing the graphs of real and synthetic data. Histogram that displays the distribution of values for each variable was plotted for synthetic data and compared with real data.

Table 9: Comparative study of Accuracy

Algorithm	Imbalanced Data	GAN	SMOTE
Logistic Regression	75	59	56
Decision Tree Classifier	66	66	63
Random Forest Classifier	75	74	58

Table 10: Comparative study of Recall (in %)

Algorithm	Imbalanced Data	GAN	SMOTE
Logistic Regression	4	62	74
Decision Tree Classifier	35	43	49
Random Forest Classifier	11	24	65

Table 11: Comparative study of ROC-AUC Score

Algorithm	Imbalanced Data	GAN	SMOTE
Logistic Regression	0.69	0.66	0.67
Decision Tree Classifier	0.56	0.63	0.61
Random Forest Classifier	0.71	0.70	0.66

7 Conclusion and Future Work

This study successfully solved the class imbalance problem in binary classification problem of loan default by generating synthetic data. Two state-of-the art techniques were evaluated. Techniques used to generate synthetic data were GAN and SMOTE and a detailed comparative analysis on effectiveness of this techniques was performed. Models were built on imbalanced and balanced data and compared on the basis of accuracy and recall. Since the objective was to build a classifier that predicts defaults and minimize loss for the lender, recall has been given priority over other metrics. Using GAN, the recall achieved by classifiers such as Logistic Regression, Decision Tree and RandomForest is 62%, 43% and 24% respectively. Similarly, using SMOTE the recall achieved by classifiers such as Logistic Regression, Decision Trees and RandomForest is 74%, 49% and 65% respectively. Overall, as a synthetic data generation technique, SMOTE outperforms GAN in terms of recall and logistic regression, delivering the highest recall score of 74%.

In the future, the performance of SMOTE can be optimized by tuning its parameter such as number of nearest neighbor for generating the synthetic sample. In the same way, TGAN can be optimized to generate synthetic data and to improve machine learning models' capacity for generalization and further lead to better accuracy and recall. Apart from visualisations, several other techniques for evaluating the quality of generated synthetic data can be implemented.

References

Abd Elrahman, S. M. and Abraham, A. (2013). A review of class imbalance problem, *Journal of Network and Innovative Computing* **1**(2013): 332–340.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J. and Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets, *IEEE Access* **8**: 201173–201198.

Ashrapov, I. (2020). Tabular gans for uneven distribution, *arXiv preprint arXiv:2010.00638* .

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

Figueira, A. and Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and gans, *Mathematics* **10**(15): 2733.

García, V., Sánchez, J. S. and Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, *Knowledge-Based Systems* **25**(1): 13–21.

Gonog, L. and Zhou, Y. (2019). A review: generative adversarial networks, *2019 14th IEEE conference on industrial electronics and applications (ICIEA)*, IEEE, pp. 505–510.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations, *International journal of data mining & knowledge management process* **5**(2): 1.

Longadge, R. and Dongre, S. (2013). Class imbalance problem in data mining review, *arXiv preprint arXiv:1305.1707* .

Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A. and Padma, V. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification, *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, pp. 729–735.

Ramyachitra, D. and Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review, *International Journal of Computing and Business Research (IJCBR)* **5**(4): 1–29.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan, *Advances in Neural Information Processing Systems* **32**.

Xu, L. and Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks, *arXiv preprint arXiv:1811.11264* .

Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science* **162**: 503–513.