Statistical Analysis of Rare Events

Indian Statistical Institute, Kolkata

Supervisor : Dr. Isha Dewan

Name : Kaulik Poddar

Roll No. : MD2207

Date of Submission : 17.05.2024

Abstract

Prediction of rare events involves identifying and forecasting events that occur with a low probability. Logistic regression, which is widely used to estimate the probability of default, or any tree based model used for classification often suffers from the problem of separation when the event of interest is rare and consequently leads to poor predictive performance of the minority class in small samples. In these types of scenarios, accuracy often misleads us. Sensitivity and precision could be a better choice. We try some penalized regression models like Firth's method, FLAC, ridge, lasso and also some tree based models like decision tree, bagging and random forest. We implement a data balancing technique **SMOTE** and again fit those models to look for improvements. We also work with more rare datasets and see how these models perform when the imbalance nature increases in the dataset. And finally we will see how important this data balancing technique is in the case of rare event data.

---

1 Introduction :

An event is a specific incident that happens at a particular time and place. A rare event is an event that has a low probability of occurring. Rare events are subset of events since they happen less frequently than regular events. In everyday life, rare events can be surprising, special, or unexpected and they capture our attention because we do not see or experience all the time these type of event. We divide our approach of this analysis into four parts :

i. Rare event data (here basically we will talk about rare events)

ii. Data processing approaches (in this section we perform EDA of our dataset, we clean our dataset and remodify the dataset)

iii. Algorithm level techniques (here we fit different models and then talk about the possible improvements)

iv. Evaluation approaches (finally we will compare the different models based on different metrics).

2 Literature Review :

i. Ozge Sezgin (2006) applied different machine learning tools on a data on credit rating system to classify whether an individual is defaulter or not.

ii. King and Zeng (2017) in their paper mentioned that logistic regression can sharply underestimate the probability of rare events. They have suggested to add some correction factor to the probability of the rare event to solve this problem.

iii. Rahman and Sultana (2017) performed two penalized logistic regression methods (Firth's method and log F prior) on some real life datasets which could be used as an improvement over logistic regression.

iv. Ogundimu (2019) in his paper mentioned about some data balancing techniques like random oversampling examples (ROSE) and synthetic minority oversampling technique methods (SMOTE) and performed a simulation study where the results indicated that SMOTE improved the outputs.

v. Andreas Hild (2021) also worked on estimating and evaluating the probability of default.

vi. Apart from them Nitesh V. Chawla et. al. (2002) in their paper mentioned that how useful the SMOTE technique is when the dataset has imbalance nature.

3 Rare Event Data :

In this section we will briefly discuss different aspects of rare event data.

3.1 Levels of Rarity :

It is quite natural there are "Levels of Rarity" in rare event data. In any domain, the rarity of events is inversely correlated with the maturity of that industry. In the early days of aviation, airplane crashes were relatively common due to the experimental nature of the technology and the lack of comprehensive safety regulations and standards. As the aerospace industry matured over time, advancements in engineering, technology, and safety protocols were made. These advancements have significantly reduced the occurrence of events like plane crashes. Today, airplane crashes are relatively rare events compared to the early days of aviation. This decrease in the rarity of plane crashes correlates with the maturity of the aerospace industry, characterized

because of well-established best practices in aircraft design, manufacturing, and maintenance. At the same time, rarity is correlated with event frequency of occurrence. In the aerospace industry example, as the industry matured, the rarity of airplane crashes decreased, which means the frequency of such events decreased over time. This correlation demonstrates how the development of the industry, along with advancements in technology and safety protocols, has reduced the likelihood of fatal events occurring, such as plane crashes. The levels of rarity are categorized into four groups below in the Figure-1.

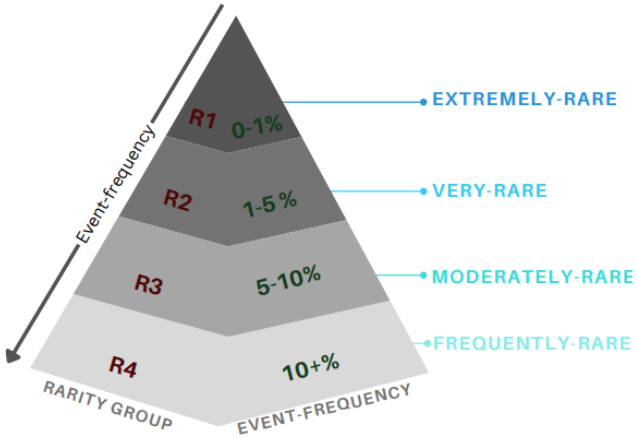

Figure 1: Levels of Rarity

3.2 Types of Rare Event Dataset :

In real-life, rare events can be observed in various domains, including medical diagnosis, fraud detection, and natural disaster prediction. We can divide the source of rare event data into eight sectors : economy, healthcare, transportation, telecommunications, manufacturing, energy, earth science, and others.

There are four types of rare event datasets : naturally rare event datasets, derived datasets, simulated datasets, and synthetic datasets.

i. **Naturally rare event datasets (RE):** Naturally rare event datasets, often referred to as rare event datasets (RE), involve situations where the occurrence of certain events is infrequent compared to the common events.

ii. **Derived datasets (DE):** A derived dataset (DE) refers to a dataset that is created or generated from one or more existing datasets through various transformations, feature engineering, or preprocessing steps.

iii. **Simulated datasets (SI) :** A simulated dataset in the context of rare events is an artificially generated dataset that follows the characteristics and patterns of real-world rare events.

iv. **Synthetic datasets (SY):** A synthetic dataset refers to artificially generated data that closely resembles

real data but is created without relying on existing data, often to expand the labelled data available for training purposes.

One can get every type of data mentioned over in the following link : https://drive.google.com/file/d/1iA4jdbFMryiLa5ZdnXF2DeAeLNabY4dI/view?usp=drive_link

3.3 Characteristics of Rare Event Dataset :

Some key characteristics which are commonly seen in these datatets are skewed class distribution, lack of data, temporal property of rare event with respect to time, class overlap between rare event and non-rare event, uncertainty of data, high dimensionality, complexity of rare event etc. But the main problem while performing classification on the rare event datasets is dealing with the **imbalance nature** of the dataset. Many times, this problem contributes to bias while making decisions or implementing policies. Thus, it is vital to understand the factors which causes imbalance in the data (or class imbalance). These hidden biases and imbalances in data can affect the association between variables, and in many cases could represent the opposite of the actual behaviour.

Most of the time rare event data suffer from the problem of imbalance which means that one of the classes has a significant higher percentage compared to the percentage of another class. In simple words, a dataset with unequal (significantly) class distribution is defined as **imbalanced dataset**. This issue is widespread, especially in binary (or a two-class) classification problems. In such scenarios, the class which has majority instances is considered as a majority class or a negative class, and the under-represented class is viewed as a minority class or a positive class. This imbalance problem is mainly dependent on the **degree of class imbalance**. One can understand the degree of imbalance by calculating the **Imbalanced Ratio (IR)**, for which the formula is given below:

$$IR = \frac{Total\ number\ of\ negative\ class\ examples}{Total\ number\ of\ positive\ class\ examples}$$

4 Data Processing Approaches :

Here we work with a dataset on rare event. We have chosen **simulated** data from the source **economy**. A significant problem for credit scoring models which must be pointed out is the unavailability of real-world credit data. The reason is that customer's credit data is confidential in most financial institutions. That is why we have to work with a simulated data. This dataset has information about mortgage applications of

made-up clients and the goal is to estimate correctly the probability of inability of applicants to repay the debt (defaulter).

4.1 Exploratory Data Analysis :

The dataset has 32581 rows and 12 variables with 7 numeric and 5 categorical. The percentage of non-defaulters and defaulters is 78.18% and 21.82% respectively. Hence, according to figure-1 our dataset is categorized into **R4 (Frequently Rare)**. The imbalance raio is 3.58, which indicates a high imbalance nature. Table-1 presents the variable names, description, and type of data for this dataset.

Variable	Description	Type
person_age	Age	Quantitative
person_income	Annual Income	Quantitative
person_home_ownership	Home ownership	Categorical
person_emp_length	Employment length (in years)	Quantitative
loan_intent	Loan intent	Categorical
loan_grade	Loan grade	Categorical
loan_amnt	Loan amount	Quantitative
loan_int_rate	interest rate	Quantitative
loan_status	Loan status (0 is non default and 1 is default)	Categorical
loan_percent_income	Percent Income	Quantitative
person_default	Historical Default	Categorical

Table 1: Variable description of the Dataset

First we clean the data set, identify the wrong values and filter the data for the modeling part. For the variables **person_age** and **person_emp_length**, we can see that there are some extreme values that do not make sense, for example, clients have age or employment duration over 100 years. To fix this, all values above 100 years of age or employment duration were removed. There are 5 extreme values and removing 5 out of 32581 observations will not affect much our models. Now we have 32576 data points to do further analysis.

Now we check for every variable whether there is any missing observation or not. Here our response variable is loan_status which has no missing observation. Apart from this, only the following two variables have missing observations : person_emp_length and loan_int_rate. These two variables have 895 and 3115 missing observations respectively. Since it is a quite large number, removing all the missing data points would not be a feasible solution. So we will impute these missing values. Generally, employment length of a person depends on the age of that person. So we impute a particular missing data by the mode of that corresponding age group. Similarly we did this for interest rate considering loan amount as the grouping factor.

Let us first look at the graphs of each quantitative predictors. From Figure-2 we can see that all the variables are positively skewed and age, income and emp length are highly positively skewed. And in Figure-3 the distribution of each categorical variable is shown.

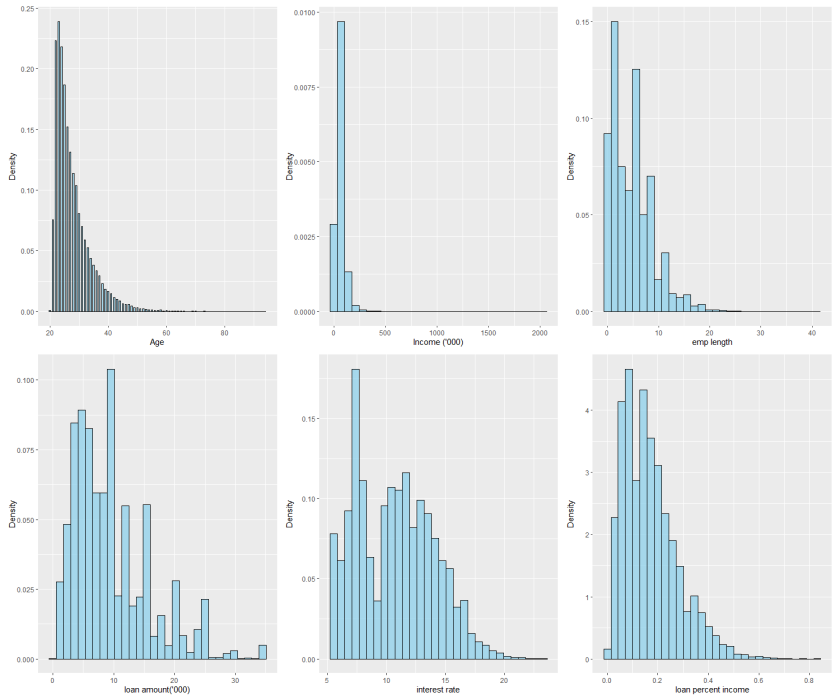

Figure 2: Distribution of Quantitative Variables

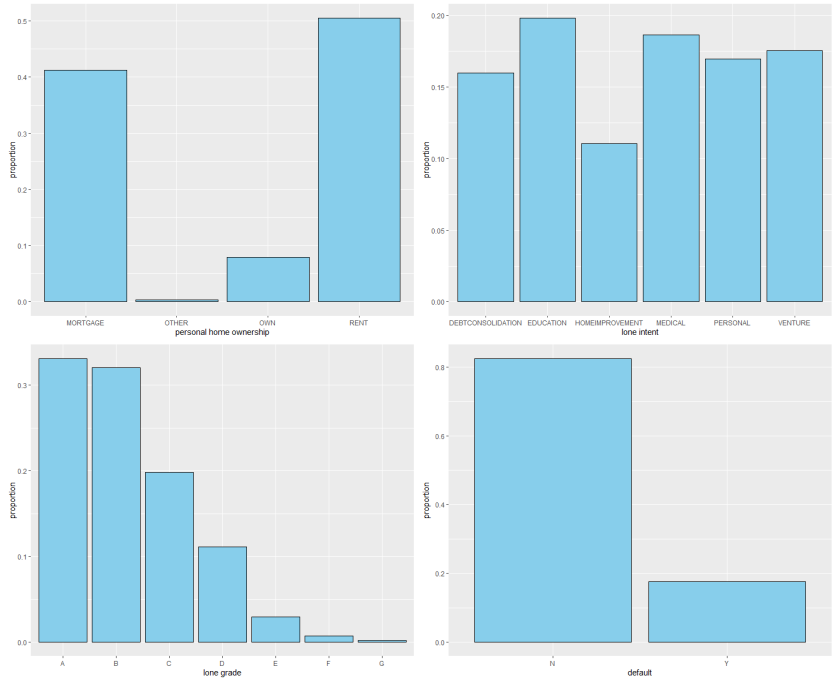

Figure 3: Distribution of Categorical Variables

Now we plot each predictor with respect to the response variable loan status and see whether the response variable depends on that predictor or not. For quantitative predictor we use boxplot and for categorical predictor we use bar diagram.

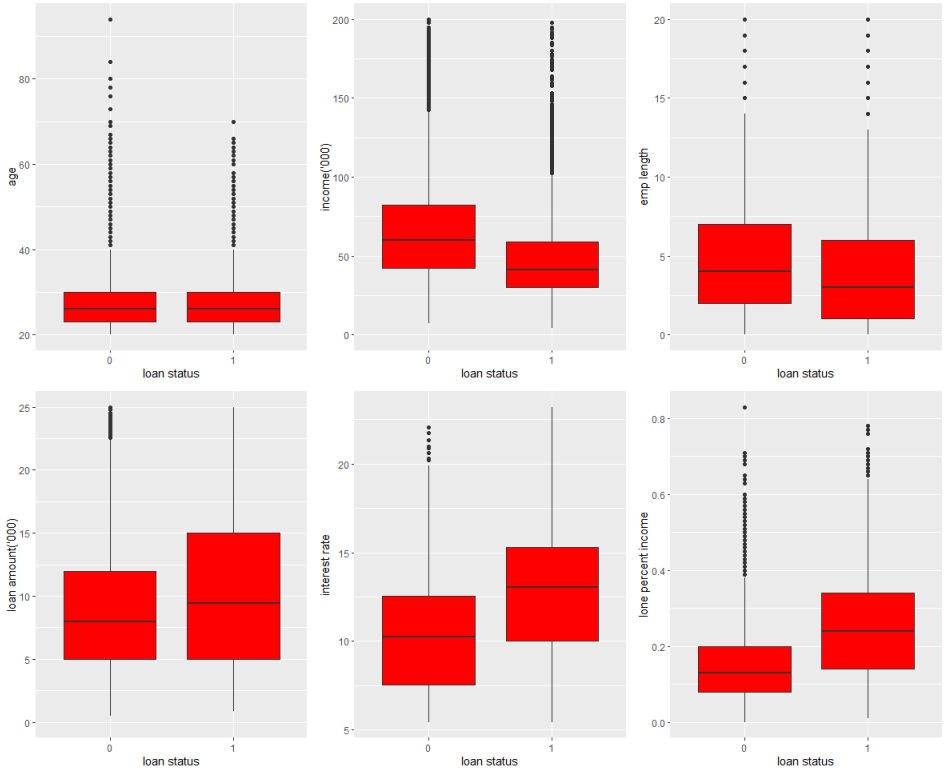

Figure 4: Boxplot of Quantitative Predictors w.r.t. loan status

From Figure-4 it seems like person_age is not a significant predictor. Then we can see that on an average the non-defaulters have more income than that of defaulters, which is quite natural. The non-defaulters have slightly higher average employment length than that of defaulters. In case of loan amount and interest rate, we see that defaulters have it higher (on an average) than non-defaulters.

Now consider Figure-5. For the predictor home ownership, the proportion of defaulter is lowest for the category **OWN** and highest for **RENT**. For the variable loan intent, we can see each category has significant effect on loan status. Then consider the variable loan grade. It is clearly visible that as loan grade moves from A to G the proportion of defaulters increases. And finally we see that if a person has a history of default then there is a high probability that he/she will be a defaulter again, which is quite natural.

Now we perform Pearson chi-square test for the independence of attributes at 0.05 level of significance. Since p-value is less than 0.05 in every case, we reject the null hypothesis that the response variable and the attributes are independent.

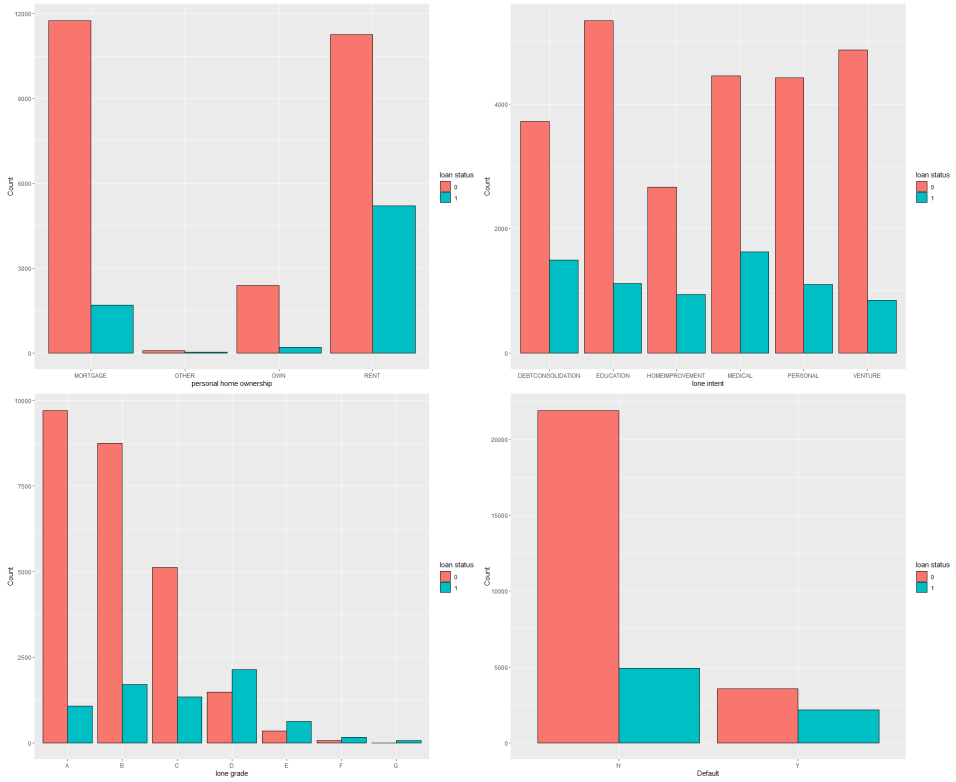

Figure 5: Barplot of Categorical Predictors w.r.t. loan status

4.2 Data Modification :

Look at the variable loan grade. It has seven categories. One may notice for loan grade A-C the proportion of defaulters is less than that of non-defaulters; whereas for loan grade D-G the proportion of defaulters is higher than that of non-defaulters. That is why we redefine this variable in a different way. We divide this variable into two categories : **Good** (for loan grade A,B,C) and **Bad** (for loan grade D,E,F,G).

There is a variable loan_percent_income which is basically proportion of loan amount and income. Since we have already used the two variables we remove this variable. There is another variable credit history length which has no significant description, so we have to omit this variable.

In our data there are four categorical predictors. Here we will use **dummification**. Dummification, also known as one-hot encoding or dummy encoding, is a process used in machine learning when dealing with categorical predictors. By converting categorical variables into binary indicator variables (0 or 1), dummification ensures that each category is treated independently. This is important because it prevents the algorithm from assuming any hierarchical relationship between the categories. Beside this, dummification allows for the handling of categorical variables with multiple categories. Each category is represented by its own binary

variable, which can then be used as input features in the model. To avoid the issue of multicollinearity (sum of dummy variables for any categorical variable is one) we remove the first dummy variable for each categorical variable.

5 Models :

After cleaning and remodifying the data we fit some models to classify our dataset. So we divide our dataset into train and test set in such away that the proportion of defaulter and non-defaulter is maintained in the train set as in original data and the train set has 80% and test set has 20% observations.

Let us first discuss briefly about the models which will be used. Consider the general set up : there are N clients with two possible events default or non-default which are governed by a Bernoulli random variable $Y_i \in \{0, 1\}$, i=1,....., N. The event of interest is Y = 1, the positive event of defaults, with associated PD, π = P(Y = 1 | X = x). The statistical problem is to estimate π = P(Y = 1 | X = x).

i. Logistic Regression : Logistic regression assumes that the probability of default, π, is a linear function of the observed covariates, i.e.

$$\pi = P(Y = 1 \mid X = x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}$$

where β is the unknown parameter. The log-likelihood function is given by

$$l(\beta) = \sum_{i=i}^{N} \left[y_i \left(\beta' x_i \right) - \log \left\{ 1 + \exp \left(\beta' x_i \right) \right\} \right]$$

ii. Firth's Method : Firth's penalized log-likelihood function for logistic regression can be written as

$$l(\beta^{Firth}) = l(\beta) + \frac{1}{2} log \mid I(\beta) \mid$$

where $\mid I(\beta) \mid$ is the determinant of the Fisher information matrix.

iii. Firth's Logistic Regression with Added Covariate (FLAC) : The objective of FLAC approach is to overcome the bias problem in average predicted probabilities of the Firth estimator. To know more about the steps of this method you may see Reference-2 (Page : 1149).

iv. Decision Tree : Decision Tree is very popular tree based machine learning algorithm which is used for both classification and regression problem. Here we apply the classification decision tree to classify whether the individual is a defaulter or not.

v. Bagging : An ensemble method is an approach that combines many simple "building ensemble block" models in order to obtain a single and potentially very powerful model. These simple building block models are sometimes known as weak learners, since they may lead to mediocre predictions on their own. The decision trees discussed generally suffer from high variance. To solve this problem we use this ensemble method "bagging" where we fit decision tree (weak learner) repeatedly to get a stable output.

vi. Random Forest : Random forest is another ensemble method. Random forests provide an improvement over bagged trees in the sense that random forest decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m = \sqrt{p}$ — that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors.

To know more about these tree based models one can see **An Introduction to Statistical Learning** (chapter 8) by Tibshirani et al. For model - i, ii and iii first we choose a threshold between 0 and 1. If P(Y=1 | X=x) is greater than that threshold, we will assign it as 1 (defaulter) and otherwise 0 (non-defaulter). Thus in model - i, ii and iii classification will be different for different thresholds. But the model - iv, v and vi directly provide us the classification of dataset.

6 SMOTE - A Data Balancing Technique :

SMOTE stands for Synthetic Minority Oversampling Technique, and is a statistical technique used to balance the class distribution of a dataset by creating synthetic minority class samples. SMOTE works by generating new instances from existing minority cases, taking samples of the feature space for each target class and its nearest neighbors, and then generating new examples that combine features of the target case with features of its neighbors.

7 Evaluation Approaches :

The first evaluation metric which comes into our mind in case of classification problem is **accuracy**. We have calculated this accuracy corresponding to every thresholds from 0 to 1 and it was seen that accuracy is maximized approximately at 0.5 threshold for the three model. But in case of imbalanced data, accuracy may

not be the best measure to compare. For example, in cancer studies the patients that have cancer represent a very small fraction of all patients. If 99% of the patients don't have cancer and 1% have, if we are trying to maximize accuracy with our model, the best option for the model would be to classify all patients as not having cancer, because it would be right in 99% of cases, but this is not acceptable. That is why **recall** and **precision** would be a better choice of metric here. Precision shows how accurate the model is for predicting positive values. Thus, it measures the accuracy of a predicted positive outcome. Recall is useful to measure the strength of a model to predict positive outcomes, and it is also known as the **Sensitivity** of a model. Both the measures provide valuable information, but the objective is to improve recall without affecting the precision. There is another metric called **area under curve (AUC)**. The curve which is talked about is **Receiver Operating Characteristic (ROC)** curve. Here we plot a graph between true positive rate (TPR) vs false positive rate (FPR) for different thresholds and then find the area under this ROC curve. The formula of the above stated metrices are given below.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	TN	FP
	Positive	FN	TP

Figure 6: Confusion Matrix for Binary Classification

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad Precision = \frac{TP}{TP+FP}$$

$$Recall(TPR) = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

8 Outputs :

8.1 Model i-iii (Without SMOTE) :

First we fit the logistic regression model. In imbalanced datasets, where one class (majority class) significantly outnumbers the other class (minority class), logistic regression tends to be biased towards the majority class. This bias can lead to poor predictive performance, especially for the minority class. That is why we will try a penalized regression model named Firth's logistic regression model. The basic idea of the firth logistic regression is to introduce a more effective score function by adding a term that diminishes the first-order term from the asymptotic expansion of the bias of the maximum likelihood estimation and the term goes to zero as the sample size increases. The beauty of this method is that it provides bias-reduction for small sample size as well as yields finite and consistent estimates even in case of separation. A major disadvantage of using

Firth's estimator for prediction in a rare events setting is that it introduces bias in predicted probabilities towards $\pi = 0.5$. This is because the determinant of the Fisher information matrix is maximized for π=0.5 and thus pushes the predicted probabilities towards 0.5 compared with the MLE. To overcome this bias problem in average predicted probabilities of the Firth estimator proposed the FLAC approach. In the Table-2 the p-values of each coefficient under each model is provided. From the outputs of these three models it is noticed

Variable	Logistic	Firth's Logistic	FLAC
intercept	0.004	0.004	0.004
person_age	0.182	0.183	0.181
person_income	<2e-16	0	0
person_emp_length	0.005	0.005	0.005
loan_amnt	<2e-16	0	0
loan_int_rate	<2e-16	0	0
person_home_ownership_OTHER	0.023	0.025	0.028
person_home_ownership_OWN	<2e-16	0	0
person_home_ownership_RENT	<2e-16	0	0
loan_intent_EDUCATION	<2e-16	0	0
loan_intent_HOMEIMPROVEMENT	0.192	0.192	0.193
loan_intent_MEDICAL	<6.98e-05	0.00007	0.00007
loan_intent_PERSONAL	<2e-16	0	0
loan_intent_VENTURE	<2e-16	0	0
loan_grade_Good	<2e-16	0	0
cb_person_default_on_file_Y	0.063	0.064	0.065

Table 2: p-values of the coefficients under each model

that the results under all the models are more or less similar. In this dataset the imbalance nature is clearly visible. For the three models it seems that age and history of default are not significant predictor at 5% level of significance. The four categories under loan_intent are significant and only one (HOMEIMPROVEMENT) is insignificant. Hence we continue with this attribute. We fit these three models again after removing the two predictors age and history of default. The rest of the predictors are coming out significant in all the three models at 0.05 level of significance.

	Predicted	
Actual	0	1
0	4890	185
1	798	642

Table 3: Confusion Matrix Under Logistic Regression Model

The confusion matrices are given in Table - 3, 4 and 5 for the first three models and we can see that they

are more or less same. Let us explain the results of one confusion matrix. Consider Table 3. The rows are indicating the actual loan status and the columns are showing the predicted loan status. Out of 1440 actual defaulters the logistic regression model (with respect to the cutoff 0.5) correctly predicts 642 number of defaulters. Similarly out of 5075 non-defaulters the model correctly predicts 4890 non-defaulters. Entries in the other confusion matrices can be interpreted similarly.

	Predicted	
Actual	0	1
0	4890	185
1	797	643

Table 4: Confusion Matrix Under Firth Logistic Regression Model

	Predicted	
Actual	0	1
0	4890	185
1	798	642

Table 5: Confusion Matrix Under FLAC

Model	Accuracy	Recall	Precision	AUC
Logistic	0.849 (0.849)	0.445 (0.448)	0.776 (0.776)	0.8616 (0.8615)
Firth's Logistic	0.849 (0.849)	0.446 (0.448)	0.776 (0.776)	0.8616 (0.8615)
FLAC	0.849 (0849)	0.445 (0.448)	0.776 (0.779)	0.8616 (0.8615)

Table 6: Evaluation Metrices Under Different Models Before and After Removing the Insignificant Predictors

The evaluation metrics are provided in Table-6. In the brackets the evaluation metrics after removing the insignificant predictors are provided. Since the confusion matrices are more or less the same and so are the evaluation metrics. Though the other metrics are quite satisfactory, recall is very poor for all of them.

8.2 Model i-iii (With SMOTE) :

We can observe from Table-6 that without SMOTE in model - i, ii and iii except recall all the other metrics give us quite satisfactory results. But for all of the models recall is very low. As we previously discussed in imbalance data it is important to have a high recall is. So we need to improve recall of these models.

The main reason behind this poor recall is the imbalance nature of the data. So we need to impose some balancing technique and we choose **SMOTE** for this. We use SMOTE on our dataset and now the percentage of defaulter is 45.57%.

After balancing the dataset we fit the three models model - i,ii and iii. In the Table-7 the p-values of each coefficient under each model is provided. From Table-7 we can see that all of the variables are significant at 5% level of significance for all of the three models.

Variable	Logistic	Firth's Logistic	FLAC
intercept	<2e-16	0	0
person_age	0.00495	0.00492	0.00489
person_income	<2e-16	0	0
person_emp_length	0.02224	0.02225	0.02220
loan_amnt	<2e-16	0	0
loan_int_rate	<2e-16	0	0
person_home_ownership_OTHER	0.00159	0.00159	0.00160
person_home_ownership_OWN	<2e-16	0	0
person_home_ownership_RENT	<2e-16	0	0
loan_intent_EDUCATION	<2e-16	0	0
loan_intent_HOMEIMPROVEMENT	0.00533	0.00534	0.00534
loan_intent_MEDICAL	0.00011	0.00010	0.00010
loan_intent_PERSONAL	<2e-16	0	0
loan_intent_VENTURE	<2e-16	0	0
loan_grade_Good	<2e-16	0	0
cb_person_default_on_file_Y	0.01273	0.01282	0.01283

Table 7: p-values of the coefficients under each model

Now we are going to classify the test dataset based on these models. Table 8,9 and 10 provide the confusion matrices with respect to the cutoff 0.5. From Table - 8,9 and 10 we can see that the confusion matrices are exactly same for logistic, firth logistic and FLAC after applying SMOTE.

	Predicted	
Actual	0	1
0	4259	834
1	1002	3262

Table 8: Confusion Matrix Under Logistic Regression Model (after SMOTE)

Table - 11 show the evaluation metrics under these models. From Table-11 and Table-6 we can see that now recall has improved significantly because of the SMOTE technique.

	Predicted	
Actual	0	1
0	4259	834
1	1002	3262

Table 9: Confusion Matrix Under Firth Logistic Regression Model (after SMOTE)

	Predicted	
Actual	0	1
0	4259	834
1	1002	3262

Table 10: Confusion Matrix Under FLAC (after SMOTE)

Model	Accuracy	Recall	Precision	AUC
Logistic	0.8037	0.7650	0.7963	0.8684
Firth's Logistic	0.8037	0.7650	0.7963	0.8684
FLAC	0.8037	0.7650	0.7963	0.8684

Table 11: Evaluation Metrices Under Different Models (after SMOTE)

8.3 Model iv-vi (Without SMOTE) :

Now we fit decision tree on the train dataset. The variables which are actually used in this tree construction are : "loan_grade_Good", "person_home_ownership_RENT", "loan_intent_MEDICAL", "person_emp_length", "person_income", "loan_amnt".

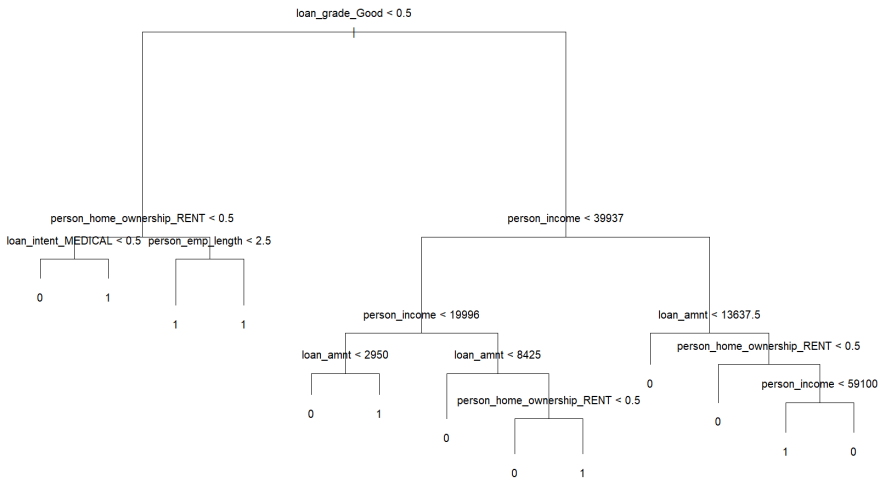

Figure 7: Decision Tree Model

Figure-7 shows the image of decision tree model and Table 12 provide the confusion matrix corresponding to this decision tree model. Now we will fit two ensemble methods bagging tree and random forest model for

	Predicted	
Actual	0	1
0	4857	218
1	527	913

Table 12: Confusion Matrix Under Decision Tree

different number of trees. We calculate misclassification error, recall and precision under the test dataset and plot them against number of trees and find the optimal number of trees should be used.

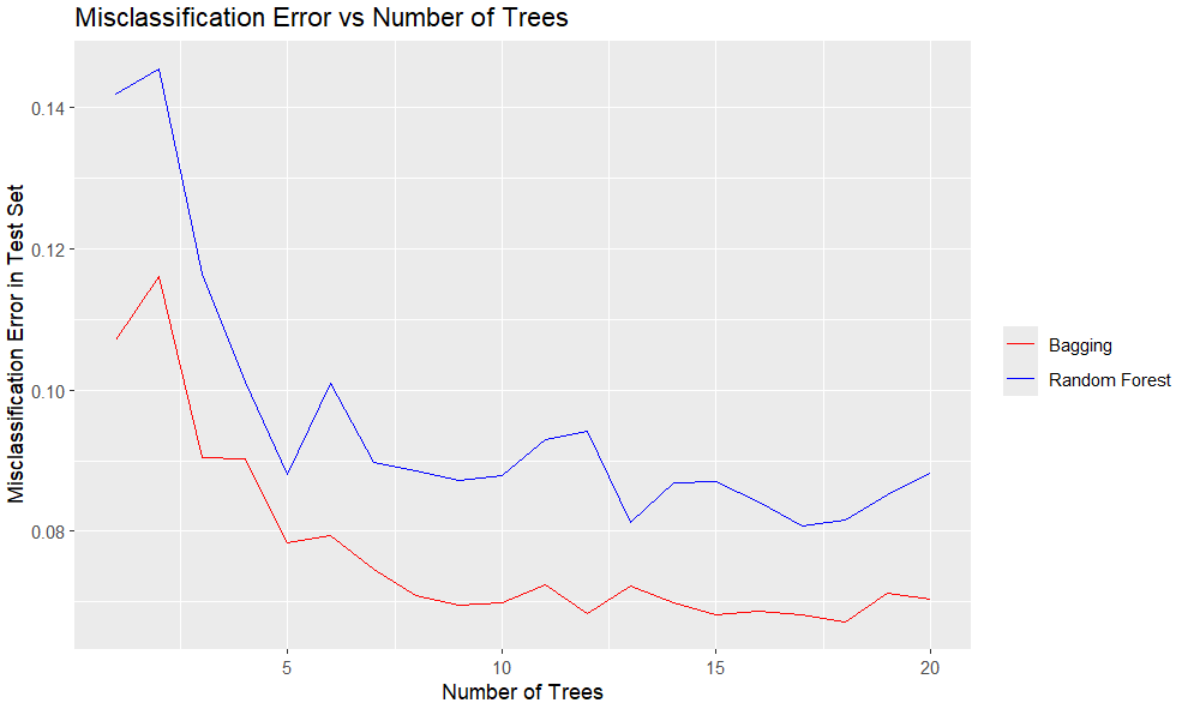

Figure 8: Misclassification Error in Test Set

So from Figure 8 and 9 we can see that misclassification error is minimum for n=17 for both techniques random forest and bagging. And sensitivity and precision are quite satisfactory at n = 17 for both models. Hence the optimal number of trees is 17. In Table 13 and 14 the confusion matrices for n=17 are given.

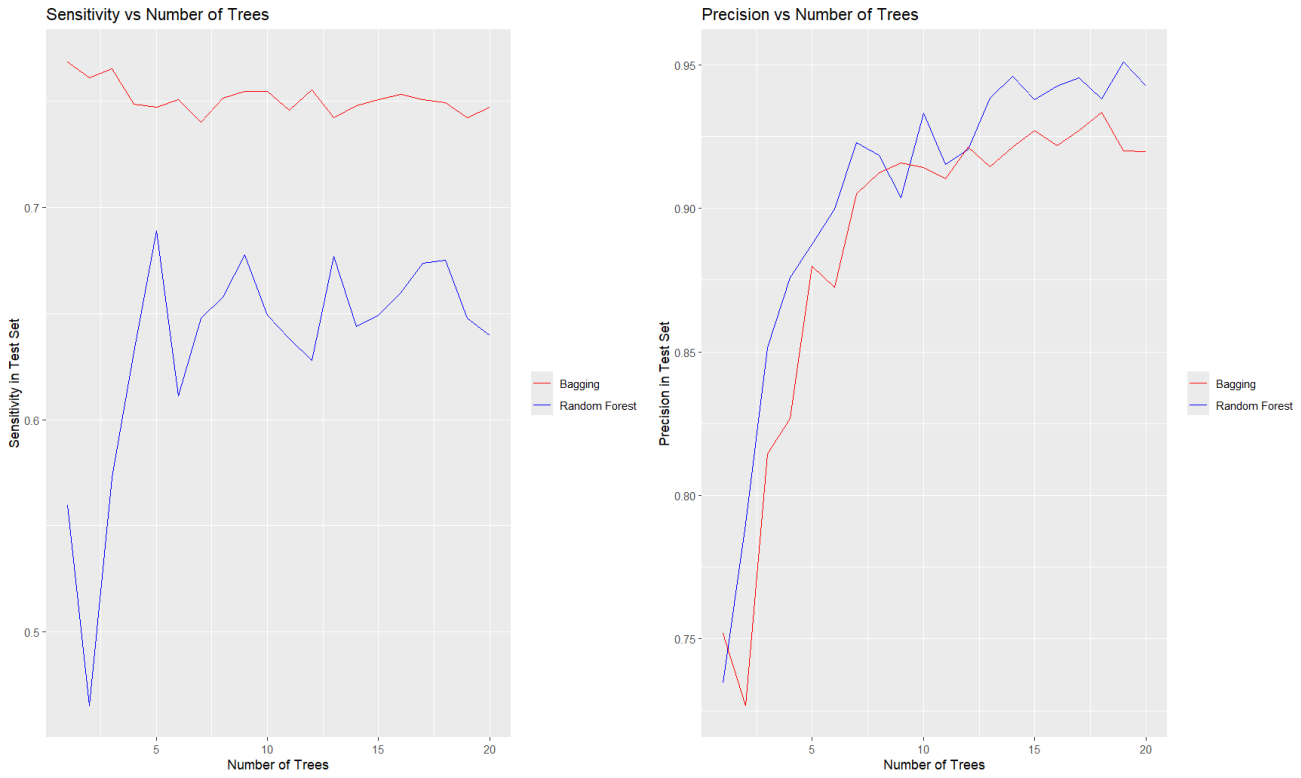

Figure 9: Sensitivity and Precision in Test Set

	Predicted	
Actual	0	1
0	5024	51
1	529	911

Table 13: Confusion Matrix for Random Forest (n=17)

	Predicted	
Actual	0	1
0	4996	79
1	361	1079

Table 14: Confusion Matrix for Bagging (n=17)

Look at the Table 15. The decision tree model indicates a good accuracy and precision and it also able to provide a recall of 0.6340. It can be improved by **bagging with number of trees equal to 17**.

Model	Accuracy	Recall	Precision
Decision Tree	0.8856	0.6340	0.8072
Random Forest	0.9109	0.6326	0.9469
Bagging	0.9324	0.7493	0.9317

Table 15: Evaluation Metrics Under Different Models

8.4 Model iv - vi (With SMOTE) :

Let's see how SMOTE will perform. And from Table-16 we can see that recall has improved significantly for all of the tree based models after implementing SMOTE.

Model	Accuracy	Recall	Precision
Decision Tree	0.8698	0.8229	0.8834
Random Forest	0.9360	0.8766	0.9811
Bagging	0.9449	0.9043	0.9730

Table 16: Evaluation Metrics Under Tree Based Models After SMOTE (n=19)

9 Tranforming Dataset in R3 Group :

Our original dataset belongs to the rarity group - R4 (Frequently Rare). Now we want to see how this models will behave if our dataset has more rarity that means having higher imbalance nature. In our original dataset the percentage of rare event was 21.82%. Now we want to convert this dataset such that the percentage of this rare event becomes between 5-10%. For this we do the following procedure :

- First we take an observation from Uniform(5,10) distribution.
- Here the observation is 6.96, so we transform our dataset such that the percentage of rare event is near 6.96%.
- So we remove $7108 - [\frac{6.96*25467}{100}] = 5336$ number of observation from minority class.
- Now the percentage of defaulter is 6.50%.
- Then we divide our dataset into train (80%) and test (20%) set such that the percentage of minority class more or less same in train set.
- Then we fit the models as mentioned in section 5.

9.1 Outputs :

Model i - iii (Without SMOTE) :

First we fit logistic, firth logistic and FLAC model and see how they perform. From Table-17 we can see that "person_emp_length" and "history of default" are insignificant predictor at 5% level of significance for all of the three models.

Variable	Logistic	Firth's Logistic	FLAC
intercept	0.2459	0.2384	0.2330
person_age	0.0033	0.0029	0.0027
person_income	<2e-16	0	0
person_emp_length	0.0585	0.0586	0.0575
loan_amnt	<2e-16	0	0
loan_int_rate	<3.35e-08	2.5067e-08	2.42478e-08
person_home_ownership_OTHER	0.5563	0.4486	0.5467
person_home_ownership_OWN	<9.63e-12	2.364775e-14	1.787459e-14
person_home_ownership_RENT	<2e-16	0	0
loan_intent_EDUCATION	<2e-16	0	0
loan_intent_HOMEIMPROVEMENT	0.7654	0.7602	0.7650
loan_intent_MEDICAL	0.0014	0.0014	0.0015
loan_intent_PERSONAL	<2e-16	0	0
loan_intent_VENTURE	<2e-16	0	0
loan_grade_Good	<2e-16	0	0
cb_person_default_on_file_Y	0.4188	0.4136	0.4210

Table 17: p-values of the coefficients under each model (R3)

	Predicted	
Actual	0	1
0	5066	29
1	273	79

Table 18: Confusion Matrix for Logistic Regression (R3)

	Predicted	
Actual	0	1
0	5066	29
1	272	80

Table 19: Confusion Matrix for Firth Logistic (R3)

	Predicted	
Actual	0	1
0	5068	27
1	274	78

Table 20: Confusion Matrix for FLAC (R3)

Model	Accuracy	Recall	Precision	AUC
Logistic	0.9445	0.2244	0.7314	0.8732
Firth's Logistic	0.9447	0.2272	0.7339	0.8733
FLAC	0.9447	0.2215	0.7428	0.8732

Table 21: Evaluation Metrices Under Different Models (R3)

Look at Table 21. The three models are very poor with respect to **recall**, which we need to improve. So again we use the technique **SMOTE**.

Model i - iii (With SMOTE) :

In this set up the data has higher imbalance nature than the original data. And the effect of this imbalance nature can be viewed in Table-21. So we need to balance the data using SMOTE technique. Hence we implement SMOTE technique on our dataset and the percentage of defaulter is now 49.34%. We follow the same procedure as it was done before. We are going to fit logistic, firth's logistic and FLAC and we see that only "person_emp_length is insignificant for all of the three models at 5% level of significance. Here we only give the table of evaluation metrics just to see the improvement of results.

Model	Accuracy	Recall	Precision	AUC
Logistic	0.8021	0.8020	0.7980	0.8776
Firth's Logistic	0.8021	0.8020	0.7980	0.8766
FLAC	0.8021	0.8020	0.7980	0.8776

Table 22: Evaluation Metrices Under Different Models (R3 and after SMOTE)

From Table-22 we can see a significant improvement in recall for all of the models. Accuracy decreases after SMOTE but the most important thing is we have gained a significant improvement in recall, which is very important metric in this data. Thus SMOTE is a very useful technique.

Model iv-vi (Without SMOTE) :

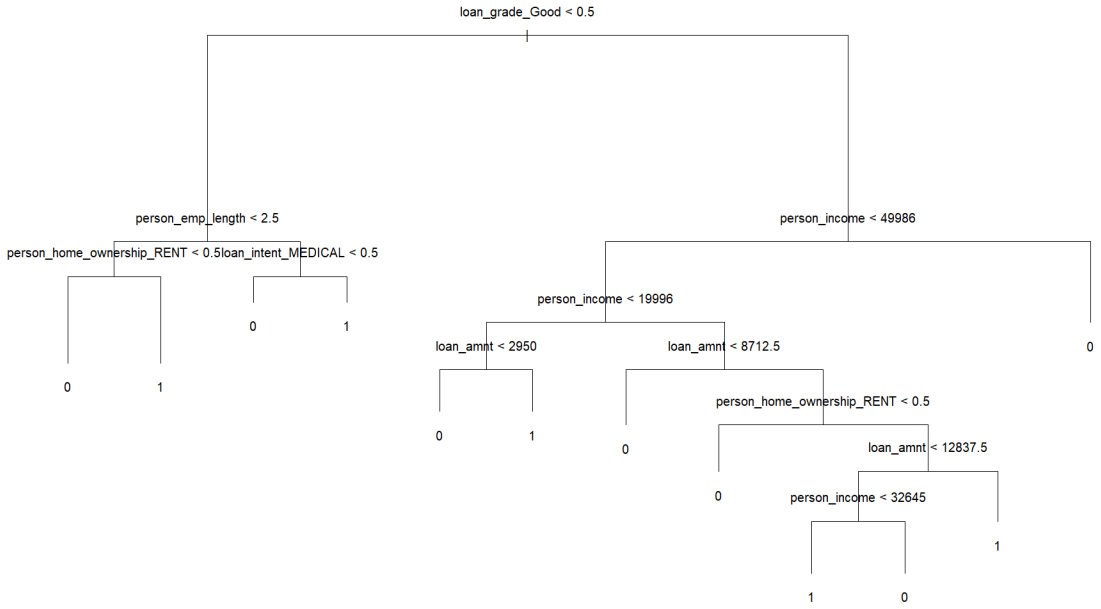

Figure 10: Decision Tree Model (R3)

Figure-10 shows the image of decision tree model and Table 23 provide the confusion matrix corresponding to this decision tree model.

	Predicted	
Actual	0	1
0	5076	24
1	192	155

Table 23: Confusion Matrix Under Decision Tree(R3)

Now we are going to check how bagging and random forest perform here. As usual first we have to find the optimal number of trees. For that we first plot misclassification error, recall and precision of test set with respect to number of trees. Our optimal number of tree should be chosen in such a way that accuracy, recall and precision in test set are as high as possible.

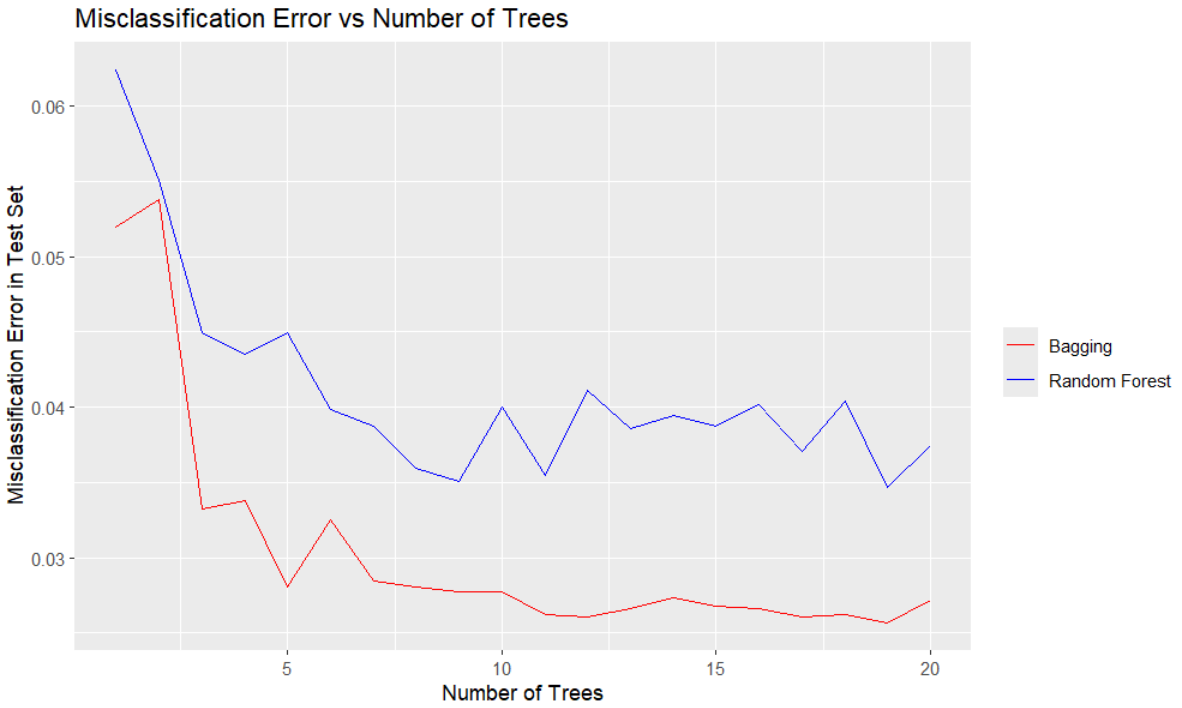

Figure 11: Misclassification Error in Test Set (R3)

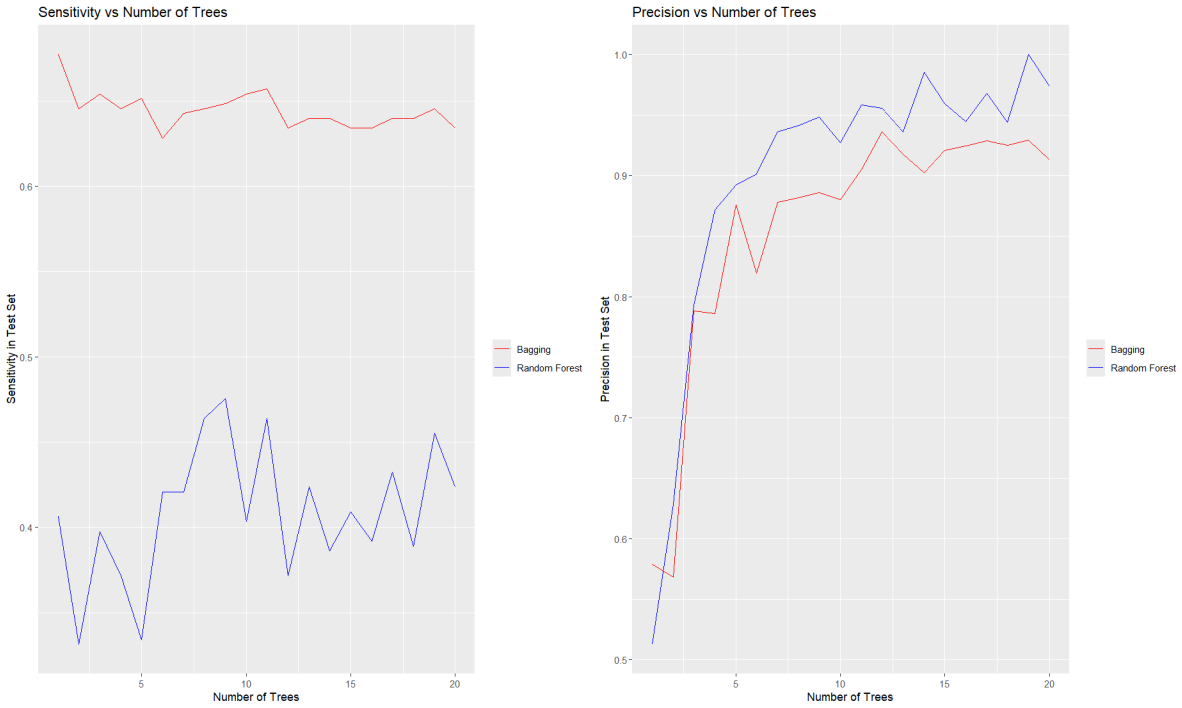

Figure 12: Sensitivity and Precision in Test Set (R3)

From Figures 11 and 12 we choose our optimal number of trees as 19. Now for n=19 we provide the confusion matrices and evaluation metrics.

	Predicted	
Actual	0	1
0	5093	7
1	204	143

Table 24: Confusion Matrix for Random Forest (n=19 and R3)

	Predicted	
Actual	0	1
0	5082	18
1	125	222

Table 25: Confusion Matrix for Bagging (n=19 and R3)

Model	Accuracy	Recall	Precision
Decision Tree	0.9603	0.4466	0.8659
Random Forest	0.9612	0.4121	0.9533
Bagging	0.9735	0.6397	0.9250

Table 26: Evaluation Metrics Under Tree Based Models Models (R3)

From Table-26 we can see that the ensemble methods are unable to achieve a satisfactory recall, so we are going for SMOTE.

Model iv-vi (With SMOTE) :

Here the optimal number of trees is 17. The evaluation metrics are given in table 27.

Model	Accuracy	Recall	Precision
Decision Tree	0.9114	0.8703	0.9458
Random Forest	0.9788	0.9590	0.9979
Bagging	0.9841	0.9762	0.9916

Table 27: Evaluation Metrics Under Tree Based Models After SMOTE (n=17 and R3)

10 Transforming Dataset in R2 Group :

Here we want to transform the dataset in such a way such that the percentage of this rare event is between 1-5%. For this we do the following procedure :

- First we take an observation from Uniform(1,5) distribution.

- Here the observation is 3.31, so we transform our dataset such that the percentage of rare event is near 3.31%.
- So we remove $7108 - [\frac{3.31*25467}{100}] = 6266$ number of observation from minority class.
- Now the percentage of defaulter is 3.20%.
- Repeat the procedure as in section 8.

10.1 Outputs :

Model i-iii (Without SMOTE) :

Only "person_age" and "history of default" are insignificant predictor at 5% level of significance for all of the first three models.

Model	Accuracy	Recall	Precision	AUC
Logistic	0.9671	0.0824	0.7142	0.8103
Firth's Logistic	0.9671	0.0824	0.7142	0.8103
FLAC	0.9671	0.0824	0.7142	0.8103

Table 28: Evaluation Metrics Under Different Models (R2)

Model i-iii (With SMOTE) :

Model	Accuracy	Recall	Precision	AUC
Logistic	0.7988	0.7925	0.8012	0.8667
Firth's Logistic	0.7989	0.7925	0.8014	0.8667
FLAC	0.7989	0.7925	0.8014	0.8667

Table 29: Evaluation Metrics Under Different Models after SMOTE (R2)

So we can see a significant improvement after implementing SMOTE.

Model iv-vi (Without SMOTE) :

Here the optimal number of trees for the ensemble methods is 20.

Model	Accuracy	Recall	Precision
Decision Tree	0.9768	0.4293	0.7835
Random Forest	0.9756	0.2768	0.9259
Bagging	0.9844	0.5649	0.9523

Table 30: Evaluation Metrics Under Tree Based Models (n=20 and R2)

Model iv-vi (With SMOTE) :

Here the optimal number of trees for the ensemble methods is 17. Here we also can see a significant improvement after implementing SMOTE.

Model	Accuracy	Recall	Precision
Decision Tree	0.9324	0.9226	0.9406
Random Forest	0.9879	0.9772	0.9985
Bagging	0.9917	0.9885	0.9948

Table 31: Evaluation Metrics Under Tree Based Models after SMOTE (n=17 and R2)

11 Evaluation Metrics vs Levels of Rarity :

Till now we have compared the evaluation metrics of different models at a particular level of rarity. Now we are going to see how these metrics behave w.r.t the levels of rarity of a particular model. We basically will see that for metrics accuracy and recall.

From Table-32 we can see that as the rarity increases i.e. as we move from R4 to R1, accuracy increases. It is because that as the rarity increases the number of actual defaulter decreases and hence there will be less number of points to be misclassified. But look at Table-34 where as rarity increases recall decreases drastically. But after implementing SMOTE in all levels of rarity the metrics provide quite satisfactory results.

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.8490	0.8490	0.8490	0.8856	0.9109	0.9324
R3	0.9445	0.9447	0.9447	0.9603	0.9612	0.9735
R2	0.9671	0.9671	0.9671	0.9768	0.9765	0.9844

Table 32: Accuracy Under Different Models (Without SMOTE)

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.8037	0.8037	0.8037	0.8698	0.9360	0.9449
R3	0.8021	0.8021	0.8021	0.9114	0.9788	0.9841
R2	0.7988	0.7989	0.7989	0.9324	0.9879	0.9917

Table 33: Accuracy Under Different Models (With SMOTE)

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.4450	0.4460	0.4450	0.6340	0.6326	0.7493
R3	0.2244	0.2272	0.2215	0.4466	0.4121	0.6397
R2	0.0824	0.0824	0.0824	0.4293	0.2768	0.5649

Table 34: Recall Under Different Models (Without SMOTE)

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.7650	0.7650	0.7650	0.8229	0.8766	0.9043
R3	0.8020	0.8020	0.8020	0.8703	0.9590	0.9762
R2	0.7925	0.7925	0.7925	0.9226	0.9772	0.9885

Table 35: Recall Under Different Models (With SMOTE)

12 Conclusion and Future Work :

In this study basically we discuss rare events and the associated challenges in the real life datasets. We tried some modifications over logistic regression. But as we have seen from the outputs the modified models did not perform well. We also tried lasso and ridge regression but they also failed to provide any improvement. So we tried some machine learning algorithms like decision tree, bagging and random forest. Though the ensemble methods are able to explain the data better than the previous methods, it is not enough. As the rarity increases **sensitivity** under each model decreases. So, all these methods are affected by data imbalance nature. So we balance the data using **SMOTE** and then again refit those models. After implementing SMOTE, all of the methods provide quite satisfactory results. We also tried k-NN and boosting algorithm, and the pattern of the result is more or less same. Now notice one thing carefully. In case of logistic, firth's logistic, FLAC and decision tree after implementing SMOTE though the recall has increased significantly but accuracy has decreased. But for the ensemble methods i.e. random forest and bagging both accuracy and recall has increased after SMOTE. Hence, these ensemble methods are best to use here. And among these two ensemble methods **bagging** provides best results in terms of all of the metrics. We also can try other machine learning algorithms to see how they perform. There are also some other data balancing techniques which can be explored.

13 References :

i. Shyalika C., Wickramarachchi R., Sheth A. (2023). A comprehensive survey on rare event prediction. (arXiv:2309.11356v1 [cs.AI] 20 Sep 2023).

ii. Gonsalves E.B.S. (2022). Different approaches of machine learning models in credit risk. NOVA Infor-

mation Management School.

iii. Hild A. (2021). Estimating and evaluating the probability of default - a machine learning approach. A thesis submitted to the Department of Statistics in Uppsala University.

iv. Kulkarni A., Batarseh F.A., Chong D. (2020). Foundations of data imbalance and solutions for a data democracy. Data Democracy, Chapter 5, Pages 83-106.

v. Cho H., Kim H., Ryu D. (2020). Corporate default predictions using machine learning.

vi. Ogundimu E.O. (2019). Prediction of default probability by using statistical models for rare events. Journal Royal Statistical Society, A, 182, 1143-1162.

vii. Puhr R., Heinze G., Nold M., Lusa L., Geroldinger A. (2017) Firth's logistic regression with rare events: accurate effect estimates and predictions? Statist. Med., 36, 2302–2317.

viii. Greenland S., Mansournia M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Statistics in Medicine, 34 3133–3143.

ix. Wang X. (2014). Firth logistic regression for rare variant association tests. Statistical Genetics and Methodology, Volume-5.

x. Volk M. (2013). Estimating probability of default and comparing it to credit rating classification by banks. Economic and Business Review, Vol 14, No 4.

xi. Vaishali G. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4.

xii. Sezgin O. (2006). Statistical Methods in Credit Rating. A thesis submitted to the Graduate School of Applied Mathematics of the Middle East Technical University.

xiii. Chen C., Liaw A., Breiman L., et al (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12):24.

xiv. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002). SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

xv. James G., Witten D., Hastie T., Tibshirani R., Jonathon T. An introduction to statistial learning.

xvi. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning.