Statistical Analysis of Rare Events

Supervised by Dr. Isha Dewan

Kaulik Poddar

Indian Statistical Institute, Kolkata

2024-05-22

Motivation :

- Analyzing rare events is important because they can cause big problems even if they happen rarely.
- By studying them, we can learn how to be ready and make plans to handle them better. This helps us avoid big losses, follow rules, and find new ways to solve problems.
- There are so many models like logistic regression model, tree based machine learning algorithm which could be used for classification.
- Because of imbalance nature of rare event data these models provide poor results. Here we will try to provide some improvements over these models.

Rare Event :

- A rare event is an event that has a low frequency of occurring. Rare events are subset of events since they happen less frequently than regular events.
- The following figure shows the different levels of rarity.

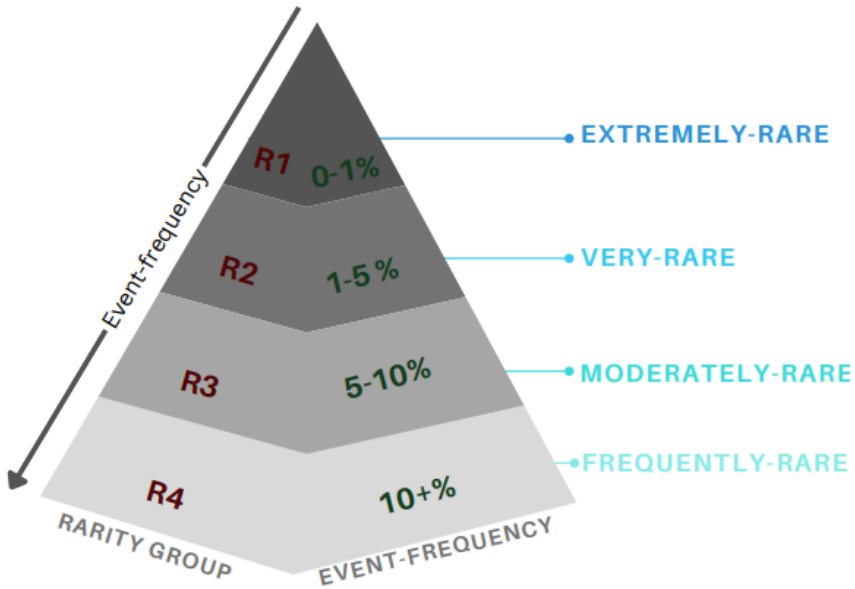

Figure 1 : Levels of Rarity

Description of Data :

- The dataset is available at https://www.kaggle.com/datasets/laotse/credit-risk-dataset?select=credit_risk_dataset.csv .
- It is simulated data.
- **Goal** : This dataset has information about mortgage applications of made-up clients. The goal is to classify correctly the applicants who are unable to repay the debt (defaulter).

Description of Data (Contd.) :

- The dataset has 32581 observations and 12 variables with 7 numeric and 5 categorical.
- The percentage of non-defaulters and defaulters is 78.18% and 21.82% respectively. The dataset is categorized into R4 (Frequently Rare).
- Variable description of the 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

Models :

- In the last presentation we talked abut the following three models :
  i. logistic regression model
  ii. firth's method
  iii. firth's logistic regression with added covariate (FLAC)

Evaluation Approaches :

- The metrics accuracy, recall and precision were used.
- Consider the following confusion matrix for binary classification:

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

Table 2 : Confusion Matrix for Binary Classification

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

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

Evaluation Approaches (Contd.) :

- We saw that accuracy often misleads us in case of imbalance data and our objective is to improve recall without affecting the precision.
- Consider another metric **Area Under Curve (AUC)** for logistic, firth's logistic and FLAC model. Here we plot Receiver Operating Characteristic (ROC) curve between true positive rate (TPR) vs false positive rate (FPR) for different thresholds and then find the area under this ROC curve.

Outputs :

Model i-iii (Without SMOTE) :

- Divide 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.
- p-values of coefficients under each model is provided below :

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 3 : p-values of the coefficients under each model

Outputs (Contd.) :

- The results under the models are more or less similar.
- It seems that age and history of default are not significant predictor at 0.05 level of significance.
- The four categories under loan intent are significant and only one (HOMEIMPROVEMENT) is insignificant. Hence the variable loan intent is significant.
- 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.

Outputs (Contd.) :

- Evaluation metrics under different models before and after removing (in the brackets) the insignificant predictors :

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 4 : Evaluation Metrics Under Different Models

- The evaluation metrics are more or less the same.
- Though the other metrics are quite satisfactory, recall is very poor for all of them.

Tree Based Models :

- We use some machine learning algorithms which are very useful technique for classification :

  iv. **Decision Tree :** Decision Tree is very popular tree based machine learning algorithm which is used for both classification and regression problem.

  v. **Bagging :** Bagging is an ensemble method that combines many simple "building ensemble block" models (here it is decision tree model) in order to obtain a single and potentially very powerful model.

  vi. **Random Forest :** Random forest is another ensemble method that provides 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 ($\approx \sqrt{p}$) 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.

SMOTE - A Data Balancing Technique :

- SMOTE stands for Synthetic Minority Oversampling Technique.
- It 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.

Outputs :

Model i-iii (With SMOTE) :

- We had seen recall is not good for all of the models.
- This might be because of data imbalance nature.
- We implement SMOTE and the percentage of defaulter is 45.57%.
- p-values of coefficients under each model is provided below :

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 5 : p-values of the coefficients under each model

Outputs (Contd.) :

- Recall has improved significantly for all of three models.
- Accuracy has decreased a little bit because of data balancing.
- Precision and AUC has increased with a small amount.
- Hence, SMOTE is a useful technique here.

Outputs (Contd.) :

Model iv-vi :

- The variables which are actually used in the decision tree model construction are :
  - loan grade Good
  - person home ownership RENT
  - loan intent MEDICAL
  - person employment length
  - person income
  - loan amount

Outputs (Contd.) :

- Want to fit bagging trees and random forest model. For that we need the optimal number of trees to be used.

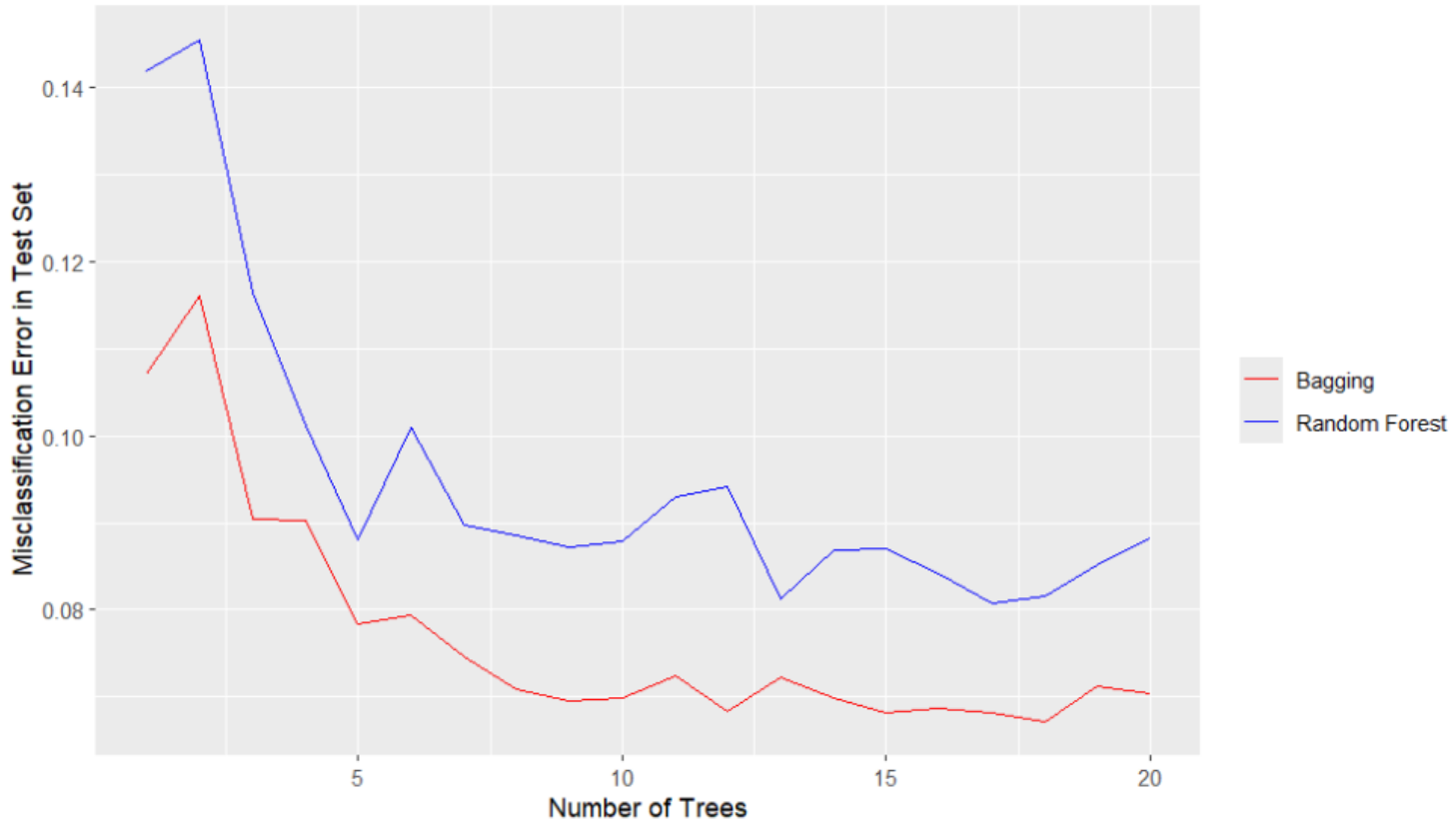

Figure 2 : Miclassification Error vs Number of Trees

Outputs (Contd.) :

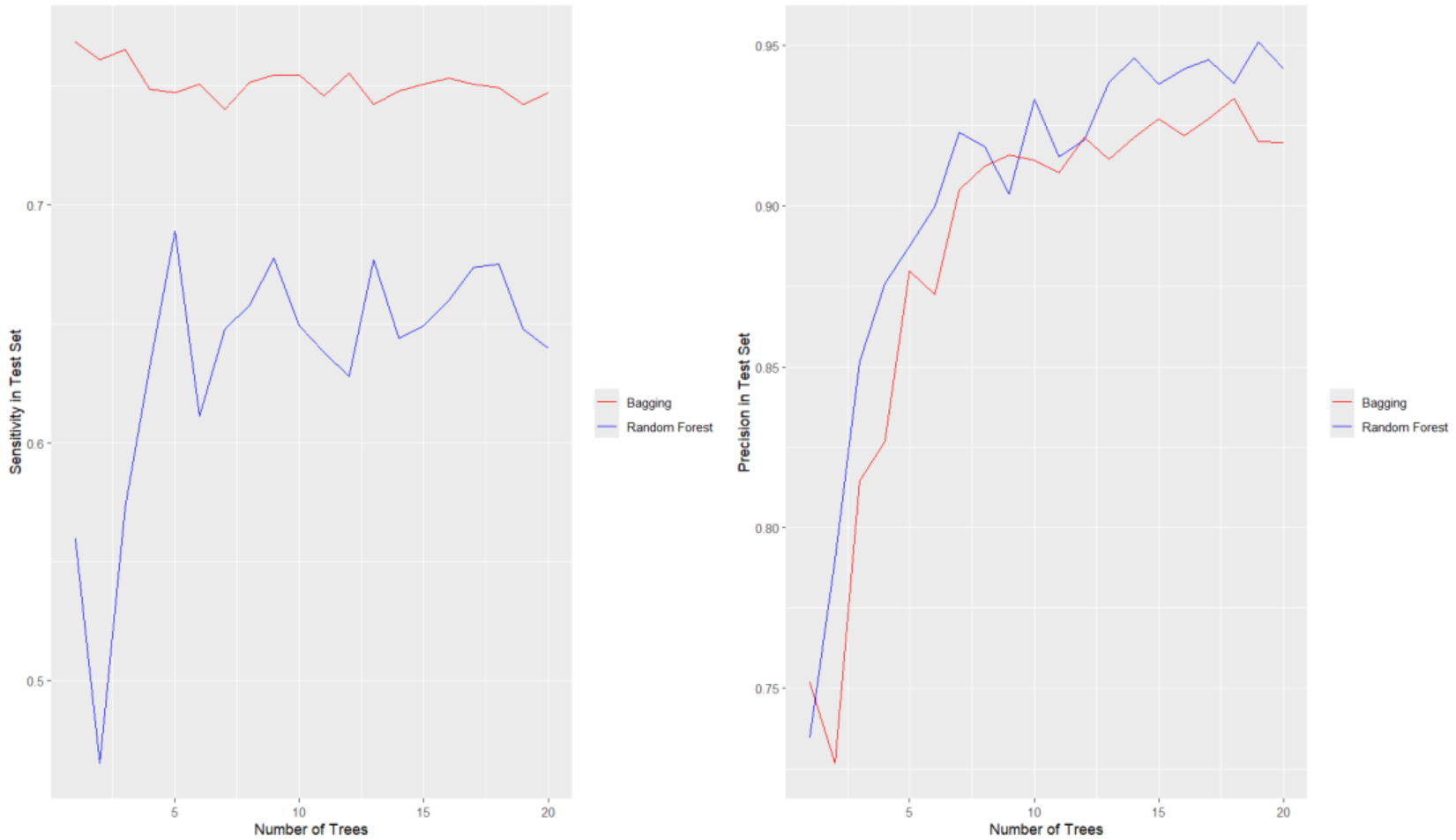

Figure 3 : Recall and Precision vs Number of Trees

Outputs (Contd.) :

- The optimal number of tree is n = 17.
- Recall can be improved to some extent by using a bagging tree with n = 17.
- After implementing SMOTE The optimal number of trees for the two ensemble methods is n = 19.
- Recall has improved significantly for all models than after implementing SMOTE.

Transforming Dataset in R3 Group :

- Want to transform the data in such a way such that the percentage of this rare event is between 5-10%.
- For this we do the following procedure :
  - First take an observation from Uniform (5,10) distribution.
  - It is 6.96, so we transform our data 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.
- We divide our dataset into train (80%) and test (20%) set such that the percentage of minority class more or less the same as in train set in the original data.
- Fit the models mentioned before.

Transforming Dataset in R2 Group :

- Want to transform the data in such a way such that the percentage of this rare event is between 1-5%.
- Follow the same procedure as in the previous just taking an observation from Uniform (1,5) distribution.

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.

Accuracy vs Levels of Rarity :

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 6 : 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 7 : Accuracy Under Different Models (With SMOTE)

- 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.

Recall vs Levels of Rarity :

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 8 : 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 9 : Recall Under Different Models (With SMOTE)

- As rarity increases recall decreases drastically.

Precision vs Levels of Rarity :

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.7760	0.7760	0.7760	0.8072	0.9469	0.9317
R3	0.7314	0.7339	0.7428	0.8659	0.9533	0.9250
R2	0.7142	0.7142	0.7142	0.7835	0.9259	0.9523

Table 10 : Precision Under Different Models (Without SMOTE)

Levels of Rarity	Logistic	Firth's Logistic	FLAC	Decision Tree	Random Forest	Bagging
R4	0.7963	0.7963	0.7963	0.8834	0.9811	0.9730
R3	0.7980	0.7980	0.7980	0.9458	0.9979	0.9916
R2	0.8012	0.8014	0.8014	0.9406	0.9985	0.9948

Table 11 : Precision Under Different Models (With SMOTE)

- As rarity increases precision decreases slowly.

AUC vs Levels of Rarity :

Levels of Rarity	Logistic	Firth's Logistic	FLAC
R4	0.8616 (0.8684)	0.8616 (0.8684)	0.8616 (0.8684)
R3	0.8732 (0.8776)	0.8733 (0.8776)	0.8732 (0.8776)
R2	0.8103 (0.8667)	0.8103 (0.8667)	0.8103 (0.8667)

Table 12 : AUC Under Different Models

- In the brackets AUC after SMOTE is provided.
- AUC is quite good in all levels of rarity.

Comments :

- Without smote, all the models perform very poor in terms of the metric recall. The bagging tree provides highest recall in all levels of rarity.
- After implementing SMOTE, recall has improved significantly for all the models. The other metrics also provide quite satisfactory results. It indicate that SMOTE is an useful technique in case of imbalanced data.
- The tree based model perform better than the logistic type models w.r.t all metrics. After implementing SMOTE the performance of two ensemble methods is outstanding in terms of all the three metrics.

Conclusion :

- We tried logistic regression model and then some penalised logistic regression model. They all performed poor in case of imbalanced data.
- Next we tried some tree based machine learning algorithm and see how they perform.
- Though the tree based models did not perform well, they performed better than the logistic type models.
- We implemented a data balancing technique called SMOTE. And we can see that after implementing SMOTE the outputs of each model improved significantly.

Conclusion (Contd.) :

- Out of these six models the two ensemble methods bagging trees and random forest are the best method to be used here in terms of the given evaluation metrics.
- We tried k-NN and boosting algorithm, and the pattern of the result is more or less same.
- We also can try other machine learning algorithms like SVM (Support Vector Machine) to see how they perform.
- There are also some other data balancing techniques like ROSE (Random Over-Sampling Examples) which can be explored.


Thank
you