Comparison of Random Forest and Logistic Regression in Classification

Hyein Jeong*

Update on August 23, 2022

Abstract

This study examines the performance of parametric and non-parametric methods. Logistic regression is used as an example of parametric methods and random forest is selected as a non-parametric model for classification. The performance of each model is evaluated by three measurement: False Positive Rate, True Positive Rate and Area Under the ROC Curve. The simulation data is manipulated to produce 8 different test sample. Across samples, logistic regression outperformed random forest. As increase the number of sample size, the performance gap measured by FPR decreased. AUC shows dominant performance of logistic regression more obviously. The largest AUC values of random forest was smaller than least AUC value of logistic model. It might suggest that the well-known disadvantage of random forest which is not suitable for data satisfying parametric test.

*Master Student at University of Bonn, email: s6hyjeon@uni-bonn.de

Contents

1 Introduction

In economic research, regression is the most commonly used method to understand the relationship between independent variables and dependent variable. Regression enables to see counterfactual situations and helps to calculate the substantial impact of a variables of interest on a social phenomenon. However, its credibility is depending on whether the assumption on the distribution of independent variables are satisfied. This characteristics underlies across parametric methods. To complement this limitation, non parametric method is postulated which does not assume distributions of each variables. Nevertheless, it becomes less powerful when the parametric test is valid, and it makes difficult to say which method is powerful than the others. Therefore, this paper will examine the performance of parametric and non parametric method with several samples.

The simulation data set comes from Bogan and Fernandez (2017). The original paper studied the impact of children with mental disability on household's investment decision. The investment decision was represented by a binary variable of 1 if the household has a certain type of asset. As a main analysis method, *Logistic Regression* was used. Of non parametric methods, this paper selected *Random Forest* to compare classification performance with logistic model. Random forest is a tree based model. A regression tree selects the most efficient split point in order and as repeats the splits, it reaches terminal node where there is no more split point. Random forest is a bundle of such regression trees. Based on bootstrap sample which allows sampling of replacement, the outcome of random forest is calculated by average of outcome of trees.

The evaluation was conducted based on simulation data set up according to the data of the original paper(Bogan and Fernandez, 2017) and the simulation data is used to produce 8 samples. The first group is baseline sample which is 70% of observations of simulation data. The second group is samples with different number of explanatory variables by their importance. The importance of variable is determined by mean decrease in Gini coefficient indicating how each variable contributes to the homogeneity of nodes and leaves in the random forest. The third group includes three samples with different sample size. Lastly, in the fourth group, there are two samples with new dummy variables of risk preference and of these two sample, the last sample went through removal of explanatory variables by their importance.

The measurement tool of performance referred to a paper which compared logistic regression and random forest(Kirasich and et al.,2018). There are three values were used to evaluate performance of each method. The first one is False Positive Rate(FPR) which is the portion of being *positive* in spite of being actually *negative*. Secondly, True Positive Rate(TPR) which is known as sensitivity was used for the analysis. It shows the portion of being *positive* of actual *positive* observation. With y-axis of TPR and x-axis of FPR, a graph can be induced, called ROC curve. The last measurement tool is Area Under the Curve(AUC) which is the area underneath ROC curve. The higher AUC value means higher accuracy of the model.

2 Theoretical Background

2.1 Logistic Regression

In economic studies, linear regression is the mostly widely used tool to understand the relationship between variables of interest. However, when it comes to categorical dependent variables, it has limitation to say that the resultant line represents all data points well. Therefore, by introducing a logit transformation, the power of prediction can be improved and the regression with logit transformation is called *Logistic Regression.* In this paper, we handle with binary dependent variable.

logistic regression estimate the probability of an event occurring which is the probability of Y being 1. When the probability of 1 is p(X), it can be defined as equation 1 and the resultant odds will be equation 2 where e is the base of natural logarithm and β are the parameters of the model.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1}$$

$$odds = \frac{p}{1-p} = e^{\beta_0+\beta_1 X} \tag{2}$$

Lastly, the logistic regression equation is as following equation 3 and it returns s-curve to predict the dependent variable based on independent variables.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \tag{3}$$

Since this method is fit in categorical variable, it is a useful for classification. The paper(Bogan and Fernandez, 2017) where this study collected data conducted their analysis on the relationship between household investment decision and children with mental disability with logistic regression. The major limitation of logistic regression is the assumption of linearity between the dependent variable and the independent variables. This assumption is underlying across parametric methods. Therefore, this study apply *random forest*,a non parametric method,to complement the limitation of logistic regression in classification.

2.2 Random Forest

Random forest is a machine learning methodology widely used across discipline. Random forest is a non-parametric and tree based method. Based on a bootstrap sample, it trains each tress and when the number of trees grows, predictions are resulted from average over trees. As an ensemble of regression trees, it complements over fitting issue in using single tree. Figure 1 shows an example of a regression tree with the original data set. It is generated by a simple regression on a binary dependent variable which indicates if a household has a safe asset or not.

$$WTRBONDS_i = AGEOFHEAD_i + lninc_i + EDUCHD_i \tag{4}$$

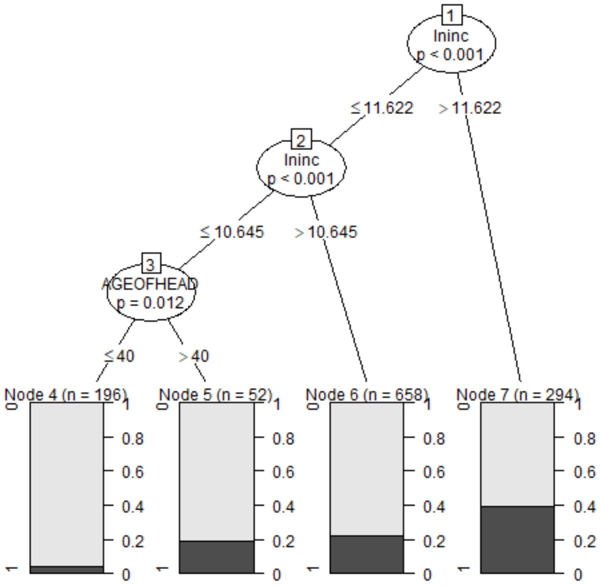


Figure 1: figure1. A regression tree example

,where *AGEOFHEAD* is the age of a household, *lninc* is a household's income and *EDUCHD* is the number of years of education of a household. This paper used `randomForest` package in `R`. the default number of tree N is 500 and `mtry` which indicates variables randomly samples as candidates at each split is set to $\lceil p/3 \rceil$ where p is the maximum number of possible direction for splitting.

To state random forest mathematically[1], random forest comprised of N randomized regression trees and the aim of this algorithm is to predict $Y \in \mathbb{R}$ with observation $\mathbf{X} \in \chi \subset \mathbb{R}$ based on a regression function $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. For each training sample has prediction functions $f_n : \chi \to \mathbb{R}$ where n is the size of train set, i.e., a train set $S_n = ((\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n))$. Of this ensemble of regression trees, $i-th$ tree has a predicted value $f_n(\mathbf{x}; \Theta_i, S_n)$ where $\Theta_1, ..., \Theta_N$ are independent random variables.

In binary classification, function $f_n$is called as a classifier which predicts $Y \in \{0, 1\}$ by following majority estimate of trees using a measurable function of $\mathbf{X}$ and S_n. In detail,

$$f_{N,n}(\mathbf{x}; \Theta_1, ..., \Theta_N, S_n) \begin{cases} 1 & \text{if } \frac{1}{M}\sum_{i=1}^{N} f_n(\mathbf{x}; \Theta_i, S_n) > 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The original paper (Bogan and Fernandez, 2017) used logistic regression to see the effects of children with mental disability on the households' investment decision which is indicated by a binary variable. In many previous studies, non parametric methods proved an improved predictive power compared to parametric methods(Hong et al., 2020). Therefore, the application of random forest for the data analyzed based on logistic regression is expected to show better prediction on household's investment decision depending on the children's mentality status. However, some says that there is a trade-off in choosing one between those methods.

[1] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

Group	Sample	Description
Group1	Baseline Sample	Baseline data with 70% observations of simulation data (n=1188)
Group2	Explanatory Sample 1	Dropped 5 explanatory variables based on the mean decrease in Gini coefficient value
	Explanatory Sample 2	Dropped 9 explanatory variables based on the mean decrease in Gini coefficient value
Group3	Small Sample	Reduced the size of baseline sample (n=832)
	Medium Sample	Increased the size of baseline sample (n=1494)
	Large Sample	Increased the size of baseline sample (n=1647)
Group 4	Risk Sample 1	Added `Risk Control` variables
	Risk Sample 2	Dropped 5 explanatory variables whose mean decrease in Gini coefficient values are the smallest from Risk Sample 1

Table 1: Type of Sample Sets

For example, while random forest might return high variability in prediction and low bias, logistic regression can result in higher bias with lower variability(Kirasich et al., 2018).

3 Data

This study benchmarked data of *Bogan and Fernandez, 2017* which investigated the impact of children with mental disability on household's investment decision. It originally from the biennial Panel Survey of Income Dynamics(PSID). It contains socioeconomic status and family structure and the authors connected it with another data, called Child Development Supplement(CDS) which includes information about children having mental- or physical disability. The data covers from 1999 to 2011 with 11763 observations.

The main analysis of the original paper was conducted with a binary dependent variable, dummy variable of children with mental disability, 27 control variables and year fixed effects. The baseline model in original paper was

$$OWN_{it} = \beta_0 + \beta_1 SpecialNeeds_{it} + \beta_k \overline{X}_{it} + \beta_h Z_{it} + \eta_t + \epsilon_{it} \tag{6}$$

where OWN is 1 if the household has a certain type of asset; *SpecialNeeds* is 1 if there is a children with mental disability; $\overline{X}$ is a matrix containing children's characteristic; Z is a matrix of household control variables. Based on this original setting, there was some manipulation of structure in regression in this paper. The main changes is that analysis was conducted only for 1999 to eliminate disadvantage resulted from using panel data in random forest and to have precise comparison its performance with logistic regression.

To explain **Data Generating Process**, the first step was generating 20 independent dummy variables. Each dummy variable was randomly produced based on the probability of 1 for each variable in the original data in 1999(n=1698). The number of children was measured by four dummy variables: `onechild, twochild, threechild` and `fourormorechild`. In this case, the variable called number of children was generated first and then it was divided into four dummy variables by one hot encoding. `Risk Control` variable which indicates groups by risk preference also went through same procedure.

Secondly, categorical variables was made with the probability for each levels in the original data. In addition, for some discrete variable and non categorical variables like `Age of Household` or `Educational years`, their values were extracted by setting same range with the one in original data. Thirdly, the only continuous variable `lninc`, a proxy of income, was resulted from a normal distribution whose mean and standard variance follows the one in original data.

This paper mainly examines performance of random forest and logistic regression on classification. For evaluation, eight types of sample was used and it can divided into four groups as described in Table 1. The first sample is 70% of the simulation data set which will be called as `Baseline sample`. The second and third sample is about removal of explanatory variables whose mean decrease in Gini coefficient is the least one. *The mean decrease in Gini coefficient* means how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The fourth to sixth sample is about the manipulation in sample size. Compared to baseline sample, the fourth sample (*Small Sample*) is less than baseline sample and the other two samples are larger than baseline sample. In Group 4, there are two samples produced by addition of explanatory variables to the baseline sample. The added explanatory variables are 6 dummy variables(*Risk Control*) of risk preference which is generated through one-hot encoding of a categorical variable. The second sample in Group 4 dropped 5 least mean decrease in Gini coefficient variables that re-calculated including risk control variables.

4 Simulation

4.1 Measurement

To evaluate the performance of random forest classification, this paper selected three criteria: false positive rate, true positive rate and area under the curve. **False Positive Rate**(FPR) is defined as the portion of being incorrectly assigned as positive but actually negative. On

Measurement	Description
False Positive Rate	$\frac{FP}{FP+TN}$
True Positive Rate	$\frac{TP}{TP+FN}$
ROC Curve	y: TPR, x: FPR
Area Under the ROC curve	the scale of area under the ROC curve

Table 2: Performance Measurement

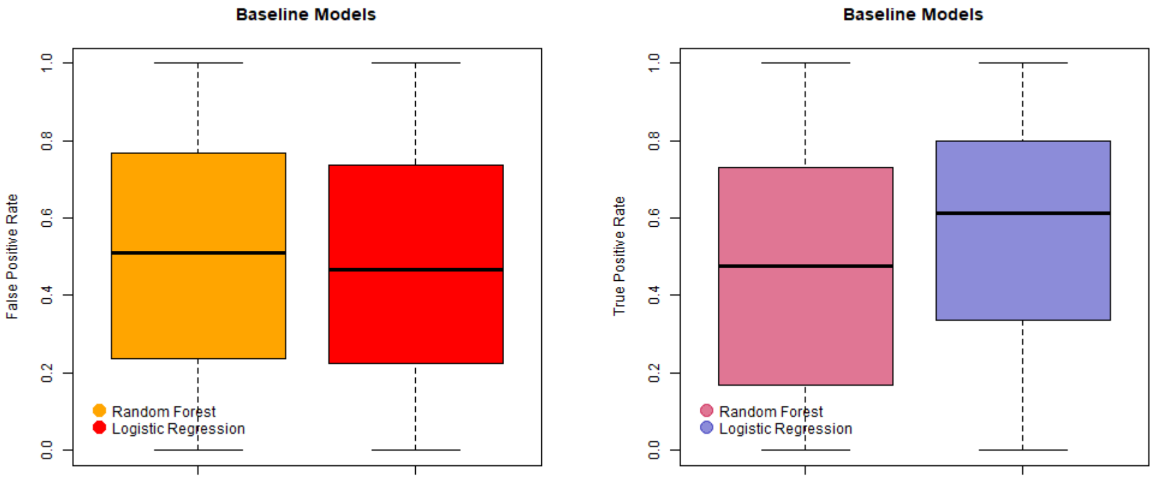


Figure 2: Group1. FPR(left) and TPR(right)

the other hand, **True Positive Rate**(TPR) which is known as sensitivity, indicates the portion of being detected as positive when it is actually positive. The curve with y-axis of TPR and x-axis of FPR is called **ROC curve** and the two-dimensional area underneath the entire ROC curve is defined as **Area Under the ROC curve**(AUC). The higher AUC means higher performance of a model because it means TPR of the model which means the ratio of conducting correct classification is more frequent than incorrect classification(FPR).

4.2 False Positive Rate and True Positive Rate

The Figure 2 shows FPR and TPR value of random forest and logistic regression with baseline sample respectively. It manifests that the logistic regression performs better than random forest when it comes to those values. In detail, logistic regression has smaller case of incorrect classification and shows remarkably better performance in correct classification than random forest.

The second group is the sample with adjustment in explanatory variables. According to the level of mean decrease in Gini Coefficient, the first sample in this group removed 5 explanatory varibles:`WTRINHERITANCE, fourormorechild, healthins, finance` and `manager`. `WTRINTERITANCE` is a dummy variable of 1 if the household has received an inheritance; `fourormorechild` is also a dummy variable of 1 if the household has four or more children; `healthins` indicates if the household has health insurance; `finance` is 1 if the household is employed in the financial sector; `manager` is 1 if the household is employed in a managerial position and zero otherwise. In the second sample in this group, four more explanatory variables are removed, which are, `regionN, unemphd, onechild` and `threechild`. `RegionN` is a proxy for the household locating in North region and `unemphd` is 1 if the household is unemployed. The result in Figure 3 shows that logistic regression showed better performance across the both samples. However, as the explanatory variables

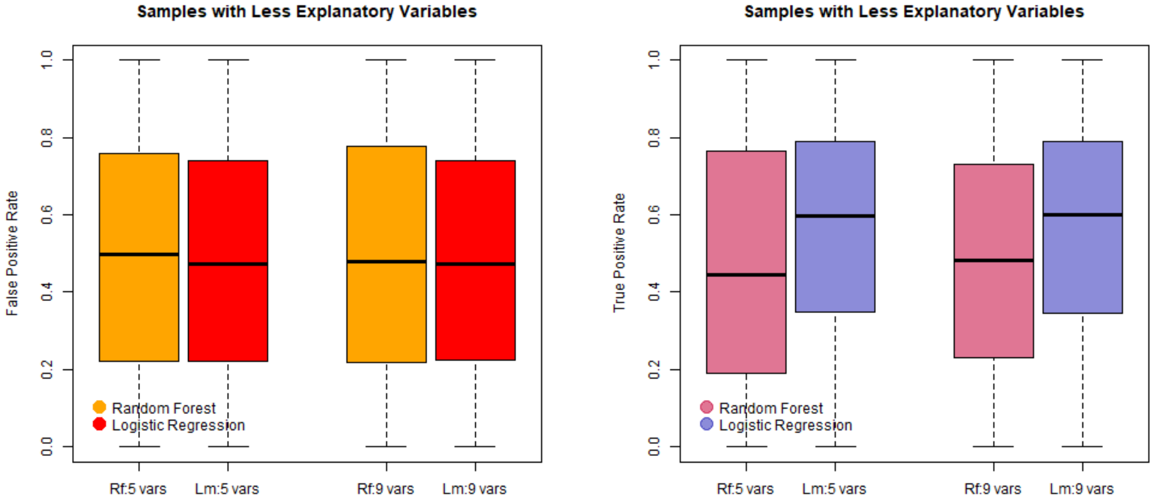


Figure 3: Group2. FPR(left) and TPR(right)

are removed, the difference in performance measurement decreased, in particularly, FPR value in the second sample returns similar value.

The third group consists of the samples with different sample size. FPR values were remained between 0.4 to 0.6 across samples and models and logistic model shows lower FPR over the samples which means it reached superior performance. Regaring to TPR, while resultant mean value of random forest model remained between 0.3 to 0.5, logistic model caused outcome between 0.5 to 0.6 which dominated random forest model. Therefore, in this group also showed that logistic model performed better but as the sample size increases, the performance gap was alleviated.

In the fourth group, new control variables were added which are dummy variable of risk preference level and the second sample of this group went through removal of 5 explanatory variables of the least mean decrease in Gini coefficient. To compare the second sample of this group with the previous sample with removal of 5 explanatory variables, the one with new control variables returns greater gap between FPR of two models. As like in the previous group, this group also manifested superior performance of logistic model over random forest.

4.3 ROC curves and AUC

Figure 6 display the comparison of performance of two models in more readable way. There is a 45 radius line across the grid which is $TPR(y) = FPR(x)$ The orange line represents random forest model and its ROC curve mostly underneath the diagonal line. This means the random forest returns more cases of false classification than correct classification.

To see detailed comparison, AUC values are suggested in Table 3. The higher AUC value means greater area under the ROC curve which means there are more case included in $TPR(y) > FPR(x)$. Therefore, the higher AUC value indicates better performance of the model. The sample with the highest AUC is **Small Sample** of Group 2. It returns

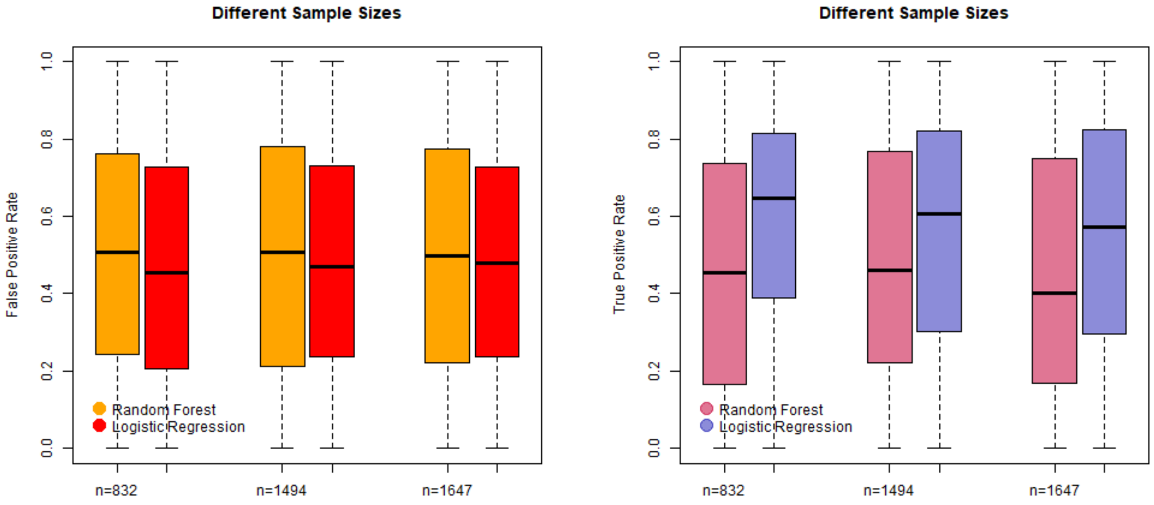


Figure 4: Group3. FPR(left) and TPR(right)

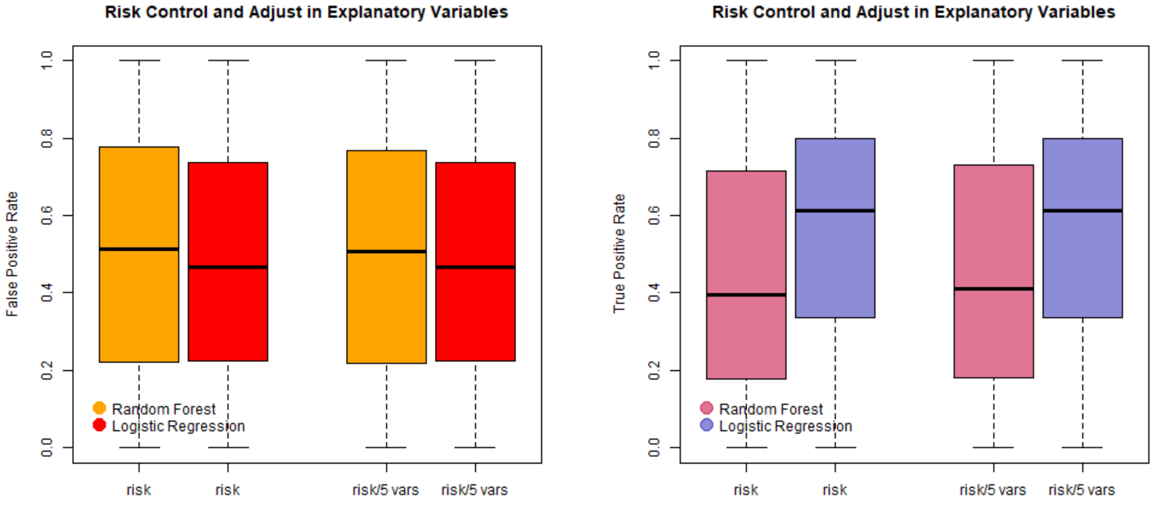


Figure 5: Group4. FPR(left) and TPR(right)

Group	Sample	Model	AUC
Group1	Baseline Sample	Random Forest	0.4562074
		Logistic Regression	0.5966265
Group2	Explanatory Sample 1	Random Forest	0.4521052
		Logistic Regression	0.5889017
	Explanatory Sample 2	Random Forest	0.5024365
		Logistic Regression	0.5863805
Group3	Small Sample	Random Forest	0.4532764
		Logistic Regression	0.6241219
	Medium Sample	Random Forest	0.4819982
		Logistic Regression	0.5825306
	Large Sample	Random Forest	0.4529289
		Logistic Regression	0.5713568
Group 4	Risk Sample 1	Random Forest	0.4539569
		Logistic Regression	0.5966265
	Risk Sample 2	Random Forest	0.4723596
		Logistic Regression	0.5966265

Table 3: AUC values

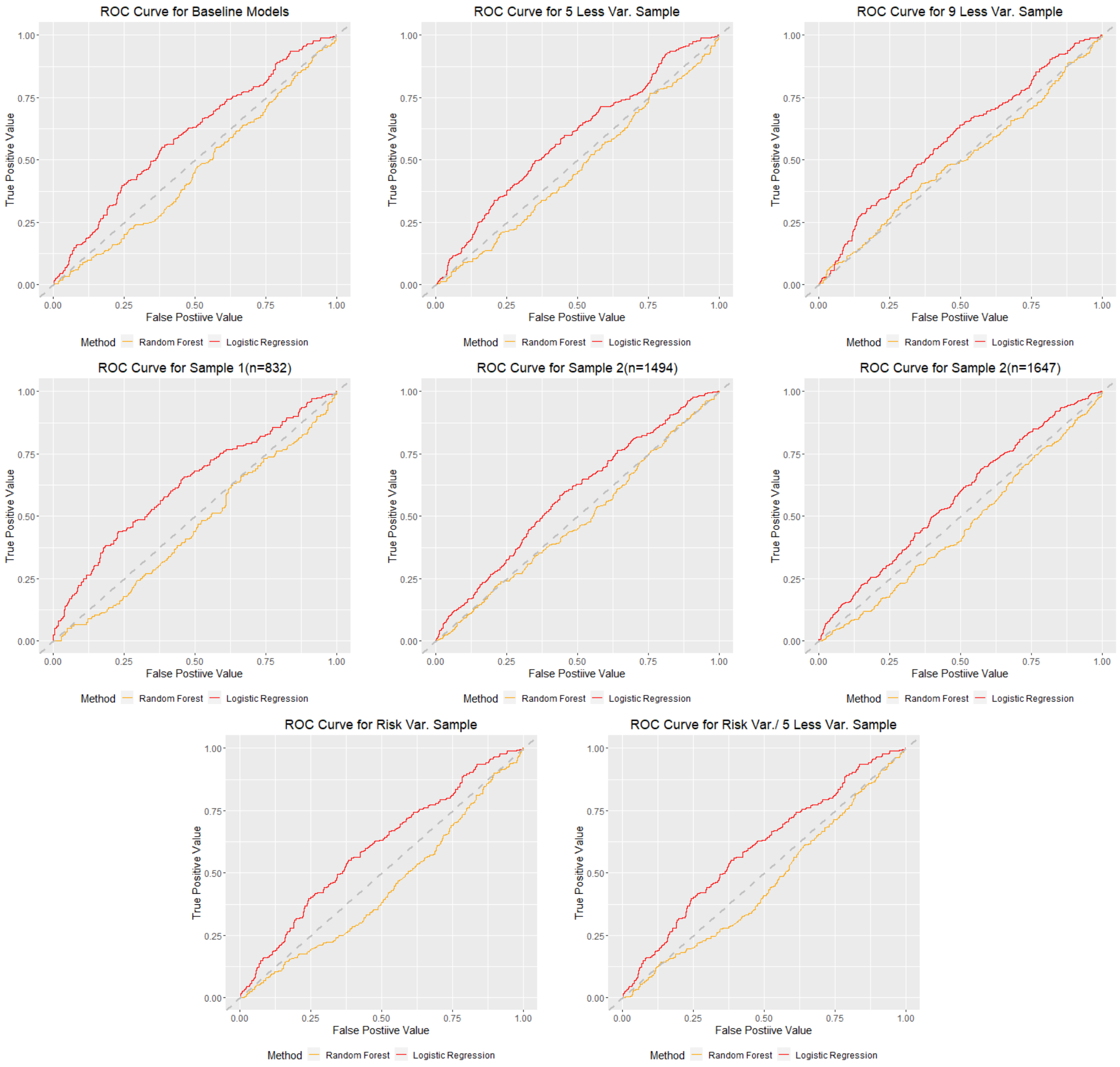
approximately 0.62 of AUC. On the other hand, the lowest AUC was found in **Explanatory Sample 1** whose AUC is approximately 0.452. Group 2 suggest that how the performance changes according to the sample size. In random forest model, the value changes 0.453, 0.456(baseline), 0.482, 0.453. Although it shows drop in AUC in the largest sample, in general it indicates better performance as the sample size grows.

5 Conclusion

This paper has its value in comparing the classification performance of two different methods: random forest and logistic regression. The measurement tools are FPR, TPR and AUC. The main analysis was conducted over 8 different samples with changes in the number of explanatory variables, sample size and addition of new variables.Across the sample sets, performance of logistic regression dominated performance of random forest regarding to both FPR and TPR. Removal of less important explanatory variables and increase in the number of sample size reduced performance gap between two models.

ROC curve has TPR as y-axis and FPR as x-axis like Figure 6 and the 45 degree line is where TPR = FPR. Therefore, if a ROC curve has more points where TPR is greater than FPR, the AUC value becomes greater. While the largest auc of random forest is 0.502 of Explanatory Sample 2 in Group 2 at Table 3, the least auc of logistic regression model is 0.571 in Large Sample in Group 3 at Table 3. It implies that logistic regression performs better than random forest regardless of sample set.

This study might suggest the well-know disadvantage of non parametric model which is not suitable for data where parametric test is valid. There is possibility that random forest cannot outperform logistic regression because each variable satisfies assumed distribution. If



ROC Curve for Baseline Models

ROC Curve for 5 Less Var. Sample

ROC Curve for 9 Less Var. Sample

ROC Curve for Sample 1(n=832)

ROC Curve for Sample 2(n=1494)

ROC Curve for Sample 2(n=1647)

ROC Curve for Risk Var. Sample

ROC Curve for Risk Var./ 5 Less Var. Sample

Figure 6: ROC Curves

so, it postulates that the performance gap with panel data might be greater than the one measured in this study. This study used a static sample with year= 1999 of panel data to control disadvantage on random forest due to dynamics. Therefore, this is open question that performance of the two models in panel data.

References

[1] Biau, G., Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[2] Bogan, V. L., Fernandez, J. M. (2017). How children with mental disabilities affect household investment decisions. American Economic Review, 107(5), 536-40.

[3] Kirasich, K., Smith, T., Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review, 1(3), 9.

[4] Janys, L. (2022). Computational Statistics. `https://github.com/LJanys/CompStat`

[5] Luo, H., Pan, X., Wang, Q., Ye, S., Qian, Y. (2019, July). Logistic regression and random forest for effective imbalanced classification. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 916-917). IEEE.

[6] Shah, K., Patel, H., Sanghvi, D., Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), 1-16.