Analysis of the Effects of Racial Discrimination on Housing Loan Supply through Machine Learning

Youhyeon Lee

January 4, 2017

Abstract

This study investigated whether racial discrimination existed in the housing loan supply market by analyzing HMDA data during 2014. Since the analysis of racial difference in the chance of getting a loan is equivalent to analyzing the treatment effect, more rigorous analytical techniques are needed so that research can be conducted in a randomized experiment situation. Therefore, the data was analyzed using the Double Machine Learning method which satisfies these conditions. In this process, machine learning techniques called LASSO and RandomForest were used. The results showed that racial discrimination occurred when using the LASSO method and that there was no racial discrimination when using Random Forest. This is because the two machine learning techniques have been created for different purposes and suggest that further analysis of this area is needed in the future. The economical contribution of this study is to suggest a new paradigm for home loan supply model by using machine learning method.

1. Introduction

Finance connects people who need money and those who want to borrow money to earn interest, which stimulates the economic vitality of individual economic entities. Housing loans also play such a role. In general, it takes a large amount of money to buy a house, and thus people apply for a mortgage on the house. The number of people who applied for housing loans only in 2014 is about 12 million. According to the United Census Bureau, the average price of new homes sold in 2014 was $345,800 and the median value was $282,800. If there are no mortgage products available, those who do not have enough money to buy a home right away will suffer from the absence of a house until they can afford.

Black, Schweitzer, and Mandell (1978) found that blacks are less likely to receive loan approval than white households, even under the same conditions, as there is a racial discrimination in home mortgage markets. Then Munnell et al. (1996) conducted a study that included variables representing the applicant's credit in that the risk of applicant's bankruptcy is an important factor in the loan approval process. Based on the result of this study, black and Hispanic applicants are still discriminated against white applicants about eight percent, even though we consider the economic credibility of individual applicant. On the other hand, Haughwout, Mayer, and Tracy (2009) used the 2/28 subprime mortgage data originated in August 2005 to estimate the racial discrimination in housing mortgage loan market, and argued that there was no racial discrimination in mortgage approval process. According to their study, the initial interest rate was rather generous and favorable for black applicants. Interestingly, however, Dell'Ariccia, Igan, and Laeven (2008) showed that the lending standard was loosened during that time, which affected the subprime mortgage crisis in 2008. They also pointed out the fact that there was racial discrimination in mortgage lending based on the result of regression using the data from 2000 to 2006.

This paper aims analyze the impact of racial discrimination on the approval of housing mortgage by randomly extracted HMDA data in 2014. In addition to the original HMDA data, information on income, unemployment rate, population, home ownership rate, the level of regional development and credit per each MSA region or each state were separately surveyed and added to the data.

Analyzing the existence and impact of racial discrimination means an analysis of the racial and ethnic treatment effects. In other words, we should divide the data into white, black, Asian, and Hispanic groups to compare the outcome of each group or the coefficient of the variable representing that group in the estimated model. However, in order to accurately estimate the treatment effect, the data should fit well into the randomized experiment. Since real economy data may not fit well in this randomized situation, this article will follow the multi-step variable selection methodology presented in Belloni, Chernozhukov, and Hansen (2012). This methodology involves several steps of regression analysis to estimate the treatment effect. First, in order to assign randomness to policy variables, we estimate a model predicting the race as a dependent variable.The estimated model depends on which machine learning technique is used, LASSO and Random Forest. Ln particular, I introduce the concept of 'Sparsity' that the only some of the independent variables are needed to predict race groups (white, black, asian, and hispanic) when using LASSO. I then estimate the models, named as m(X)s, for predicting the races of the housing loan applicants and include m(X)s as independent variables in the model for predicting approval of home loans. In estimating this second model, we reintroduce the concept of 'sparsity' that only a part of the data I have as independent variables is necessary to predict the possibility of approvals of mortgage loan. We then apply the LASSO penalty methods once again to select appropriate variables.If all of the final variables are selected by following this two-stage process, then we estimate the treatment effect by doing the logistic linear regression without LASSO penalty. When analyzing data using Random Forest, there is no concept of Sparsity and all variables are used to estimate the model. That is,

1. Select some of the variables needed to predict policy variables or estimate predictive models using all independent variables and the results were used to create artificial policy variables which have randomness - LASSO and Random Forest
2. Estimate the model needed to predict loan approval by using some or all variables including artificial variables-LASSO and Random Forest
3. Estimate the treatment effect using the model selected in Step 2

In the meanwhile, Chernozhukov et al. (2016) find that these estimated estimates are inadequate and propose a new method called "orthogonalized" or "Double ML" estimator. This is to deduce the new coefficient by removing the cause of the error in the above methodology. This study ultimately estimates the treatment effect according to the "Double ML" method.

The economic contribution of this study is as follows. Unlike previous studies with similar research themes, I can estimate the treatment effect more accurately by applying the process of granting radnomness to policy variales to estimate the treatment effect under the randomized experiment situation. This accuracy can help guide the direction of the fair-lending policy.

In addition, by introducing machine learning techniques in the course of research, it is possible to approach the housing loan problem in a completely different direction from the existing research methods. Previous studies have determined that the basic information provided by HMDA is insufficient to analyze the impact of racial discrimination. Therefore, the researchers attempted to analyze by adding to the model the variables that are considered necessary to estimate the denial probability according to their intuition and discussion.On the other hand, I will try to find out what variables are really important for estimating the denial probability within given independent variables through the LASSO method. If we can find out that other variables besides the variables that were considered important for predicting the denial probability are also meaningful variables, we will be able to develop a model for housing loan supply. Random Forest method, the other method which is used in this paper, allows the various characteristics of individual data within a given data set without collectively analyzing all data. When analyzing data in this way, we can present new perspectives on the housing loan market and future research directions by calculating different results from existing research results.

2. Data

The United States has forced banks and other financial institutions to disclose records of housing loans under the Home Mortgage Disclosure Act. In this paper, we analyze the data of 2014 to find out whether racial discrimination existed in the supply of home loans in the year. Especially, I proceeded the analysis on the data which are not guaranteed by other government agencies. This is because it is more difficult to estimate a more accurate treatment effect by analyzing data other than the conventional type, since the applicants who have been assured by an external institution are more likely to receive loan approvals themselves, regardless of their race.

Dependent variable in the data is action taken on the loan approval. There are two categories for the action taken on the loan approval, which , are divided by 0 or 1, depending on whether the applicant's application for a home loan has been approved or not. In the case of the home loan application applicant withdrawing the application by himself / herself and in case of missing documents, it can not be classified into 0 or 1. Therefore, the relevant row was excluded from the data. I eliminated the instances only about the action taken on the preapproval request because independency among data can be undermined if one applicant requested the preapproval for several financial institutions and got response from them. The independent variable contains brief information about the applicant and the residential area. Among many variables, I removed variables that would be meaningless to predict the dependent variable. Numerical variables were left unchanged and categorical variables were dummy processed. The list of Numerical, categorical and eliminated variables are shown in the table 1.

The policy / treatement variables of this study are dummy variables indicating wheter the applicant is black, asian, others, and Hispanic. Others includes all other cases except white, black, and Asian. Therefore, in this paper, there are a total of four policy / treatment variables. In this process, applicant race name variable and applicant ethnicity name variable are dummy processed. Then white variable and non - hispanic variable are removed respectively among dummy processed variables from the two categorical variables due to multi - collinearity problem. This allows us to interpret the coefficient in the way that there exists racial discrimination compared to the white if an estimated coefficient of a policy variable is bigger than zero, no discrimination if the estimated coefficient is zero and got generous loan judgement result if the estimated coefficient is less than zero. Likewise for the Hispanic line, the + sign indicates that the Hispanic line is more likely to be rejected when applying for a home loan.

In addition to the data provided by the HMDA, I collected and added variables that would help predict the dependent variable. These variables include nine variables such as unemployment rate per state or msamd, expenditure on urban construction, population, and home ownership rate. These variables and their sources are shown in table 2.

In particular, it is not perfect data for individual applicants, but I added the 2014 Vantage score for each state provided by Experian. The Vantage score is a credit score made by three major credit bureaus: Experian, Equifax and TransUnion. The Vantage score is used by many lenders as an important source of information on loan approval judgements. In addition to this, I also added 2014 Late payment, average dept for each state provided by Experian in relation to credit score. I also added variables that seemed as deeply related to home loan reliability provided by the Nerdwallet web page.

The data for 2014 have about 12 million rows, of which 25,000 data were randomly sampled at 0.2 %. The number of rows used for the actual analysis was about 8,500 as a result of removing rows with missing data and other rows not suitable for data analysis.

Table 1: A summary of variables from the HMDA data

Numerical variable	**Categorical variable**	**Eliminated variable**
tract to msamd income	property type name	purchaser type name
popluation	owner occupancy name	rate_spread
minority popluation	msamd name	state abbr
number of owner occupied units	state name	loan type name
number of 1 to 4 family units	loan purpose name	sequence number
loan amount	lien status name	respondent id
hud median family income	hoepa status name	denial reason name
applicant income	co-applicant sex name	county name
	co-applicant race name	census tract number
	co-applicant ethnicity name	year
	applicant sex name	application date indicator
	applicant race name	agency name
	applicant ethnicity name	agency abbr
	preapproval name	

Table 2: Additional variables and their sources

Additional variables	**sources**
msa average income	BEA

msa change income	BEA
msa unemployment rate	BLS
msa log population	U.S. census bureau
state home ownership rate	U.S. census bureau
state construction spending	U.S. census bureau
state obtaning lawful permanent resident status	DHS
loan to income ratio	HMDA
applicant to msa average income	HMDA, BEA
Residents who spend over 35% of their income on housing	nerdwallet, U.S. census breau
change in median income 2009 ∼ 2014	nerdwallet
serious mortgage deliquency rate	nerdwallet, CoreLogic
Foreclosure inventory	nerdwallet, CoreLogic
mortgage health score	nerdwallet
2014 Vantage score	Experian
2014 Late Payments	Experian
2014 Average Debt	Experian

In the course of the research, the categorical variables are treated as dummy variables. In this process, multiple collinearity problems arise if dummy variables are not removed for each categorical one. Therefore, the dummy variables removed from the data used for analysis are as follows.

Table 3: Eliminated variables from each categorical variable

Categorical variables	**Eliminated dummy variables**
property_type_name	Manufactured housing
owner_occupancy_name	Not owner-occupied as a principal dwelling
msamd_name	Abilene - TX
State name	AK
loan_purpose_name	Home improvement
lien_status_name	Not secured by a lien
hoepa_status_name	HOEPA loan
co_applicant_sex_name	co_app_gender_Female
co_applicant_race_name	co_app_race_White
co_applicant_ethnicity_name	co_app_ethn_Not Hispanic or Latino
applicant_sex_name	applicant_gender_Female
applicant_race_name	applicant_race_White
applicant_ethnicity_name	applicant_ethn_Not Hispanic or Latino
preapproval_name	Preapproval was not requested

3. Econometric Framework

This study attempts to analyze whether racial discrimination exists in the supply of housing loans. In other words, when we divide the applicants of the housing loan into groups according to their race, it is analyzed whether the probability of receiving loan approvals according to the affiliated group is changed. In general, when conducting research on this subject, logistic regression analysis is performed because the dependent variable, loan approval, takes 0 or 1 as its value, and the regression model is as follows.

$$Pr(loan\ denial|X) = \frac{1}{1 + exp\{-(\beta_0 + Black\theta_B + Asian\theta_A + Others\theta_O + Hispanic\theta_H + X'\beta)\}}$$

If the model is estimated afterwards, the probability that the loan approval will be rejected depends on the applicant's race and ethnicity by θ values. When we compare the outcome of each group, we estimate the treatment effect. However, when a researcher estimates the treatment effect, each data should be assigned randomly to a policy group. Unless the data is randomly assigned to each policy group, it can not be excluded that deriving the outcome difference between the groups may have other external influences as well as race. If so, the estimated treatment effect is not correct. However, the data that economic researchers are trying to analyze does not meet these randomized experiment conditions. Thus, as suggested by Belloni, Chernozhukov, and Hansen (2012), I estimate models that predict policy variables following the quasi-experimental approach which allows us regard each data as if it were randomly assigned to each group if the policy or treatment variable is represented by small but a sufficient number of other variables. Then the policy variables re-estimated by the estimated model are called policy variables with randomness. And then estimate the treatment effect by including the policy variables with randomness in an independent variable set instead of the original policy variables. The specific process of estimating the treatment effect θ, which is the target parameter of this study, depends on the researcher's estimating the treatment effect using LASSO and Random Forest.

3.1 Logistic regression with LASSO penalty

The LASSO method suggested by Tibshirani, Hastie, and Wainwright (2015) imposes a constraint that the sum of the absolute values of the coefficients of the variables, l_1 -norm, is below a certain constant in the regression analysis. Then LASSO estimates the coefficient of some variables to 0 according to the constraint and removes the corresponding variable from the model.

$$\hat{\beta}_{MLE} = arg \max_{\beta_0, \beta} -\frac{1}{N} \sum_{i=1}^{N} \{y_i(\beta_0 + x_i^{'}\beta) - \log(1 + exp\{\beta_0 + x_i^{'}\beta\})\}$$

$$subject\ to\ |\beta_0| + |\beta_1| + |\beta_2| + |\beta_3| + \cdots + |\beta_p| = \|\beta\|_1 \leq t$$

In the process of finding out the appropriate t value, cross-validation is used. If the t value is arbitrarily determined, the given data set is divided into N groups, which is usually divided into 10 equal parts. These ten sub datasets are numbered sequentially from one to ten. Then, firstly, we designate the first sub dataset as a test set and other sub datasets as training datasets. Secondly, we designate the second sub dataset as a test set and other sub datasets as training datasets. . . . Lastly, we designate the tenth sub dataset as a test set and other sub datasets as training datasets. Once the training set and test set are determined for each of the ten steps, the LASSO penalty is applied to the training sets determined at each step to find the values of $\hat{\beta}$. Let the dependent variable of the test set in step 1 is y_{1i} and the estimated coefficients from the training set is $\hat{\beta}_1$. And calculate the value of $\sum_{i=1}^{N}(y_{1i} - x_i^{'}\hat{\beta}_1)^2$. Add these calculated values in step from one to ten.

$$\sum_{i \in S1}(y_{1i} - x_i^{'}\hat{\beta}_1)^2 + \sum_{i \in S2}(y_{2i} - x_i^{'}\hat{\beta}_2)^2 + \sum_{i \in S3}(y_{3i} - x_i^{'}\hat{\beta}_3)^2 + \cdots\cdots + \sum_{i \in S10}(y_{10i} - x_i^{'}\hat{\beta}_{10})^2$$

We calculate the result values for every value of t and select a specific value of t as an optimal t^* if the constant t derives the smallest result. Although a researcher could arbitrarily set t, it was named machine learning in the sense that the machine would calculate all possible values and select the optimal value among them. Once the optimal t^* value is determined, the estimated coefficient values can be obtained. In this case, variables with estimated coefficients of 0 are classified as meaningless variables to predict dependent variables. Finally, we can predict each dependent variable with variables whose coefficients are non-zero. In this way, LASSO can be used as a model selector to select only variables that are important for predicting dependent variables among all given variables. Without a process of selecting meaningful variables using the LASSO penalty machine learning technique, the researcher can try to consider too many variables without knowing what is an important variable. Then, while more energy than necessary may be required to analyze the data, considering only the statistically significant variables using the LASSO penalty can lead to an efficient decision-making process. For example, Tibshirani, Hastie, and Wainwright (2015) applied a method of imposing a LASSO penalty on the logistic regression instead of the previously used 'Back probing 'algorithm for classifying handwritten numbers(Le Cun, Boser, Denker, Henderson, Howard,Hubbard and Jackel (1990)).

3.1.1 The progress with LASSO

This section describes the process of estimating the treatment effect by applying the LASSO penalty to regression analysis. In order to estimate a more accurate treatment effect, we first apply randomness to the policy variable. In this study, policy variables are dummy variables indicating whether applicants are black or Asian, belonging to other groups, or belonging to Hispanics. Let us denote the original dummy variables as B, A, O, and H, respectively. Then, we estimate the models that can predict B, A.O, and H, and denote each model as $m_B(Z_i), m_A(Z_i), m_O(Z_i)$, and $m_H(Z_i)$. These are expressed collectively as $m_X(Z_i)$. The model used to predict loan approval, outcome variable, is denoted $g(Z_i)$. Then, according to this estimated model, the results are again estimated and denoted as $\widehat{B}, \widehat{A}, \widehat{O}, \widehat{H}$. These are policy variables to which randomness is assigned. In estimating the models $m_X(Z_i)$ and $g(Z_i)$, which predict the policy variables, I can estimate the model adequately by using only some of the variables given from the original data. It is called Sparsity. This Sparsity allows researchers to take into account more appropriate data of a manageable size from very large data. However, the researcher does not know precisely which variables are included in some of these variables in advance. In this study, this process of variable selection depends on machine learning techniques, not on intuition of researchers or economics theory about the field. The researcher specifically uses LASSO to select the variables to be included in the set of the independent variables to predict the policy variables and the outcome variable.

In particular, with respect to selecting particular variables from among all variables, we add the condition that the model is exogenous to the unselected variables in the model, if the function forms of $m_X(z_i)$ and $g(z_i)$ are linear. That is, the larger the sample size N, the more linear the function approximates the original function. Therefore, the concrete form of $m_X(z_i)$ and $g(z_i)$ is as follows.

$$m_X(Z_i) = x_i^{'}\beta_{m_X} + r_{m_{X}i} \tag{1}$$

$$g(Z_i) = x_i^{'}\beta_{g0} + r_{gi} \tag{2}$$

This requires one more condition to the Framework that the corresponding errors $r_{m_{X}i}$와 r_{gi},which occur when approximating a linear function to the original function decrease as the sample size n increases.

$$\{\bar{E}[r_{gi}^2]\}^{1/2} \lesssim \sqrt{s/n} \;\; and \;\; \{\bar{E}[r_{m_{X}i}^2]\}^{1/2} \lesssim \sqrt{s/n}$$

Basically, I employ logistic regression to analyze data since policy variables have only 0 or 1 as their values. In estimating the coefficients in the regression analysis process, specific models are estimated by applying the LASSO penalty to select only the significant variables for each dependent variable.

$$\widehat{\beta}_B = arg\max_{\beta} -\frac{1}{N}\sum_{i=1}^{N}\{B_i(\beta_{0B} + x_i^{'}\beta_{m_B}) - log(1 + exp(\beta_0 + x_i^{'}\beta_{m_B})\} + \lambda\|\beta_B\|_1$$

$$\widehat{\beta}_A = arg\max_{\beta} -\frac{1}{N}\sum_{i=1}^{N}\{A_i(\beta_{0A} + x_i^{'}\beta_{m_A}) - log(1 + exp(\beta_0 + x_i^{'}\beta_{m_A})\} + \lambda\|\beta_A\|_1$$

$$\widehat{\beta}_O = arg\max_{\beta} -\frac{1}{N}\sum_{i=1}^{N}\{O_i(\beta_{0O} + x_i^{'}\beta_{m_O}) - log(1 + exp(\beta_0 + x_i^{'}\beta_{m_O})\} + \lambda\|\beta_O\|_1$$

$$\widehat{\beta}_H = arg\max_{\beta} -\frac{1}{N}\sum_{i=1}^{N}\{H_i(\beta_{0H} + x_i^{'}\beta_{m_H}) - log(1 + exp(\beta_{0H} + x_i^{'}\beta_{m_H})\} + \lambda\|\beta_H\|_1$$

Then, the coefficients of the insignificant variables are estimated to be 0 and removed from the model. And the individual models $\widehat{m}_X(Z_i)$ are estimated with only surviving variables, we can obtain the results according to the estimated model.

$$\widehat{B}_i = \frac{1}{1 + exp\{-\widehat{m}_B(Z_i)\}} \tag{3}$$

$$\widehat{A}_i = \frac{1}{1 + exp\{-\widehat{m}_A(Z_i)\}} \tag{4}$$

$$\widehat{O}_i = \frac{1}{1 + exp\{-\widehat{m}_O(Z_i)\}} \tag{5}$$

$$\widehat{H}_i = \frac{1}{1 + exp\{-\widehat{m}_H(Z_i)\}} \tag{6}$$

Then, secondly, when estimating the model with the loan approval as a dependent variable, these

$\widehat{B}, \widehat{A}, \widehat{O}, \widehat{H}$, are selected as some variables of the independent variables instead of the original racial and ethnicity variables because they got the same properties as those assigned randomly. At this time as well, all parameters belonging to the data given by the HMDA and the data added by the researcher may not be necessary in predicting the approval of the loan. Therefore, only variables that are significant in predicting dependent variables are selected from given data. I denote the outcome variable as Y and the variable selection process is as follows. I also employ logistic regression to analyze data for this process because the outcome variable has only 0 or 1 as its values.

$$\begin{aligned}(\widehat{\alpha}, \widehat{\beta}) = arg \max_{\alpha,\beta} -\frac{1}{N}\sum_{i=1}^{N}\{Y_i(\beta_0 + \widehat{B}_i\alpha_B + \widehat{A}_i\alpha_A + \widehat{O}_i\alpha_O + \widehat{H}_i\alpha_H + x_i^{'}\beta) \\ -log(1 + exp(\beta_0 + \widehat{B}_i\alpha_B + \widehat{A}_i\alpha_A + \widehat{O}_i\alpha_O + \widehat{H}_i\alpha_H + x_i^{'}\beta)\} + \lambda\|\beta\|_1\end{aligned}$$

Then, as the coefficients of the logistic regression analysis are estimated by applying the LASSO penalty, the coefficients of the insignificant variables are estimated to be zero and removed from the model, and only the variables that are significant in predicting the approval of the loan will ultimately survive. Then, according to the result of the logistic regression analysis with the LASSO penalty, the coefficients of the insignificant variables are estimated as 0 and removed from the model, and only the variables that are significant in predicting the approval of the loan will ultimately survive. In this way which was mentioned above, In the first step, I select the variables needed to predict the policy variables to create the artificial variables $\widehat{B}, \widehat{A}, \widehat{O}, \widehat{H}$. And in the second step, I selected the variables needed to predict the dependent variable among all the variables given in the data. In the third step, I estimate the coefficient values by performing a logistic regression without the LASSO penalty with the final selected variables. In the second step, the linear combination of the variables selected in addition to the artificial variables is expressed as $g(Z_i) = x_i^{'}\beta_{g0} + r_{gi}$.

$$\begin{aligned}(\breve{\alpha}, \breve{\beta}) = arg \max_{\alpha,\,\beta} -\frac{1}{N}\sum_{i=1}^{N}\{y_i(\widehat{B}_i\alpha_B + \widehat{A}_i\alpha_A + \widehat{O}_i\alpha_O + \widehat{H}_i\alpha_H + g(Z_i) \\ - \log(1 + exp(\widehat{B}_i\alpha_B + \widehat{A}_i\alpha_A + \widehat{O}_i\alpha_O + \widehat{H}_i\alpha_H + g(Z_i))\} \qquad (7)\end{aligned}$$

I denote the estimated outcome variable which is obtained by the estimated model $\widehat{l}_0(Z) = \widehat{B}_i\breve{\alpha}_B + \widehat{A}_i\breve{\alpha}_A + \widehat{O}_i\breve{\alpha}_O + \widehat{H}_i\breve{\alpha}_H + \widehat{g}(Z_i)$ as $\widehat{Y}$.

However, chernozhukov et al. (2016) reported that the estimators $\breve{\alpha} and \breve{\beta}$ estimated by the above method are not appropriate estimators because the linear function converges slowly to the original function. Since there are errors due to the difference between the original function that we do not know and the approximated function, they create a model that removes the cause of this error from the outcome and treatment variables and named it " Orthogonalized " or " Double ML ". Finally, I estimate the treatment effects according to this methodology. The specific model is as follows.

$$W = V_B\theta_B + V_A\theta_A + V_O\theta_O + V_H\theta_H + U$$

Here, $V_B = B - \frac{1}{1+exp(m_B(Z))}$, $V_A = A - \frac{1}{1+exp(m_A(Z))}$, $V_BO = B - \frac{1}{1+exp(m_O(Z))}$, $V_H = H - \frac{1}{1+exp(m_H(Z))}$ and $W = Y - \frac{1}{1+exp(l_0(Z)_i}$. In particular, since we estimate the model assuming its function as a linear function, we first estimate $m_X(Z_i)$ and $l_0(Z_i)$, and then estimate θ. Therefore, the estimated W and V variables are obtained first. That is, $\widehat{V}_B = B_i - \widehat{B}_i$, $\widehat{V}_A = A_i - \widehat{A}_i$, $\widehat{V}_O = O_i - \widehat{O}_i$, $\widehat{V}_H = H_i - \widehat{H}_i$ and $\widehat{W} = Y_i - \widehat{Y}_i$. Then the " orthogonalized " or " Double ML " estimator in this study is as follows.

$$\breve{\theta}_X = \Big(\frac{1}{N}\sum_{i=1}^{N}\widehat{V}_{Xi}^2\Big)^{-1}\frac{1}{N}\sum_{i=1}^{N}\widehat{V}_{Xi}\widehat{W}_i \tag{8}$$

3.2 Random Forest

Random Forest is a method of developing the classification and regression trees proposed by Breiman et al. (1984). The classification tree gathers data that have a homogeneous tendency among the data of various characteristics. A Classification tree is created by creating a child node based on the best split criterion from the root node that contains all the data. At this time, each node tries to split data according to the binary questions that can come from all the variables included in the data, and actually divides the data into two groups based on the binary question which causes the impurity to decrease the most when generating the child nodes. At this time, in selecting the best split, the Random Forest belongs to the machine learning technique in that the machine selects the best split criterion among all binary questions based on all calculated impurity reduction value at the child node compared to the parent node.

Breiman (2001) proposed Random Forest, which was developed to randomly generating multiple trees in a data classification model from only using one tree. Now, when selecting the best split criterion for each node in the process of creating one tree, we need not to consider all dichotomous criteria that can be created from all independent variables, but rather to use one, few variables or a linear combination of few variables, then select the criterion that maximizes the impurity reduction value of the child node compared to the parent node among all the dichotomies that can be obtained from them, and divides the data into two. In this way, when one tree is maximized, it is not pruned. Following this principle, we bootstrap the existing data sets to create multiple data sets and then construct a random forest model by creating a tree for each data set. Then the individual data in the terminal nodes of each tree would have been classified as either approved or not. Forest then votes the result of each tree classifier for individual data and assigns the data to the most selected group. There is upper bound of the probability that a misclassification will occur when the data is classified in this way, and the upper bound is determined by the correlation between each tree constituting the forest or the accuracy of the individual tree.

$$mr(\boldsymbol{X}, Y) = P_{\boldsymbol{\Theta}}(h_k(\boldsymbol{X}) = Y) - \max{}_{j \neq Y} P_{\boldsymbol{\Theta}}(h_k(\boldsymbol{X}) = j) \tag{9}$$

$$P_{\boldsymbol{X},Y}(mr(\boldsymbol{X}, Y) < 0) \leq \bar{\rho}(1 - s^2)/s^2 \tag{10}$$

The Random Forest does not analyze all the data collectively but gathers the data of the same nature according to the split criterion and classifies whether the data is approved or not according to the affiliated group. This allows heterogenity situations in which the data have different characteristics, enabling more detailed and accurate data analysis for researchers.

3.2.1 The progress with Random Forest

Random forest method estimates the predictive model using all variables without using the concept of sparsity when estimating the treatment effect. However, as in the case of LASSO, the treatment effect is estimated after giving randomness to the policy variables. In this case, White and Nonhispanic are added to the main variables, and they are expressed as Wh, NH. First, we estimate Random Forest models that use B, A, O, H, Wh, and NH as dependent variables to obtain the policy variables with randomness. $F_{B0}, F_{A0}, F_{O0}, F_{H0}, F_{N0}, F_{W0}$, and F_{NH0} are models that predict whether the applicant estimated according to the random forest model corresponds to each policy variable, and $F_X(Z)$ represents these models at once.

$$T_{Bi} = F_{B0}(Z_i) + \xi_{Bi}, \quad E[\xi_{Bi}|Z] = 0$$

$$T_{Wi} = F_{Wh0}(Z_i) + \xi_{Wi}, \quad E[\xi_{Whi}|Z] = 0$$

$$T_{Ai} = F_{A0}(Z_i) + \xi_{Ai}, \quad E[\xi_{Ai}|Z] = 0$$

$$T_{Oi} = F_{O0}(Z_i) + \xi_{Oi}, \quad E[\xi_{Oi}|Z] = 0$$

$$T_{Hi} = F_{H0}(Z_i) + \xi_{Hi}, \quad E[\xi_{Hi}|Z] = 0$$

$$T_{NHi} = F_{NH0}(Z_i) + \xi_{NHi}, \quad E[\xi_{NHi}|Z] = 0$$

Then, the values classified according to the above models,$T_{Bi}, T_{Ai}, T_{Oi}, T_{Hi}, T_{Wi}, T_{NHi}$, are referred to as artificial variables derived from the Random Forest model. These are policy variables that are granted ranomness through the Random Forest. After that, we re-estimate the Random Forest model that predicts whether the loan is approved. In this case, the model is estimated by putting the estimated values above, $T_{Xi}, whereX = B, A, O, H, W, NH$, not B, A, O, H, W, and HN, into the independent variables.

$$Y_{Bi} = g(T_B, Z) + \nu_{Bi}, \quad E[\nu_{Bi}|Z, T] = 0$$

$$Y_{Wi} = g(T_W, Z) + \nu_{Whi}, \quad E[\nu_{Whi}|Z, T] = 0$$

$$Y_{Ai} = g(T_A, Z) + \nu_{Ai}, \quad E[\nu_{Ai}|Z, T] = 0$$

$$Y_{Oi} = g(T_O, Z) + \nu_{Oi}, \quad E[\nu_{Oi}|Z, T] = 0$$

$$Y_{Hi} = g(T_H, Z) + \nu_{Hi}, \quad E[\nu_{Hi}|Z, T] = 0$$

$$Y_{NHi} = g(T_{NH}, Z) + \nu_{NHi}, \quad E[\nu_{NHi}|Z, T] = 0$$

The newly estimated Y_{Xi} values are the results classified as whether the applicants have been approved for loan approval or not, as information on each policy variable is given. The treatment effect is then basically estimated as follows. For example, we can find whether there is a difference between getting a loan approval between black and white as follows.

$$\theta_{BW} = E[g(T_B, Z) - g(T_{Wh}, Z)]$$

However, to estimate the treatment effect θ under conditions that satisfy the orthogonality condition as given by Belloni, Chernozhukov, and Hansen (2016), I exploit the following equation.

$$\psi_{BW}(W,\theta,\eta) := (g(T_B,Z) - g(T_{Wh},Z)) + \frac{B(Y - g(T_B,Z))}{F_B(Z)} - \frac{T_{Wh}(Y - g(Wh,Z))}{F_{Wh}(Z)} - \theta$$

$$\eta(Z) := (g(T_B,Z), g(T_{Wh},Z), F_B(Z), F_{Wh}(Z)), \quad \widehat{\eta}(Z) := (\widehat{g}(T_B,Z), \widehat{g}(T_{Wh},Z), \widehat{F}_B(Z), \widehat{F}_{Wh}(Z))$$

Then analyze the square root of θ, which makes $\frac{1}{N}\sum_{i=1}^{N}\psi_{BW}(W,\theta,\widehat{\eta}) = 0$, the difference between black and white, that is, the treatment effect. This process is repeated in the process of analyzing the difference between Asian and Caucasian, Others and Caucasian, and Hispanic and Nonhispanic.

4. Analysis Results from the logistic regression with LASSO penalty

4.1. Variable selection for the race and Hispanic policy variables

In order to analyze the existence of racial discrimination in home loan supply, this study first implemented the process of giving randomness to policy variables.Of the more than 500 variables given, If policy variables, such as race and Hispanic variables, can be linearly controled with small but a sufficient number of variables, then can treat the estimated policy variables as if the data belonging to which race and Hispanic group were randomly assigned. Therefore, this study selected variables necessary to predict race and Hispanic through the LASSO penalty. The variables selected for each race and Hispanic variable and their coefficients are as follows. Each coefficient value was truncated at the fourth decimal place. The difference between the conventional logistic regression and the logistic regression with LASSO penalty is that the variables to be discussed are already classified as meaningful variables for predicting the dependent variable by the LASSO penalty. Therefore, it is not necessary to test whether the coefficient is 0 through standard error or t-statistics and to check whether the variable is statistically significant or not. Variables that have survived are not zero coefficients.

4.1.1. Black or African American

The estimated model $\widehat{m}_B(Z)$, which predicts whether applicant is black or not by applying LASSO penalty to logistic regression, is as follows.

Table 4: Selected variables for the black group from the logistic regression with LASSO penalty

coeff	variables	coeff	variables
0.0041	tract_to_msamd_income	-0.4239	Not applicable.2

-0.0002	population	-0.6294	Secured by a first lien
0.0504	minority_population	-0.5420	Secured by a subordinate lien
0.0005	number_of_owner_occupied_units	-1.0890	co_app_gender_Male
0.0002	number_of_1_to_4_family_units	0.4212	co_app_gender_No co-applicant
-0.0006	loan_amount_000s	-0.3004	co_app_race_Asian
0.0000	hud_median_family_income	6.0252	co_app_race_Black or African American
0.0026	applicant_income_000s	0.1017	co_app_race_Information not provided by applicant
-1.1202	AZ	0.7595	co_app_race_No co-applicant
-0.7889	CA	0.4159	co_app_ethn_No co-applicant
-0.0638	IN	-1.9231	applicant_gender_Information not provided by applicant
0.4510	LA	-0.3112	applicant_gender_Male
0.6850	MA	-3.2198	applicant_ethn_Hispanic or Latino
0.0201	NC	-1.0088	applicant_ethn_Information not provided by applicant
-0.0256	NM	0.0000	msa_average_income
0.4457	NY	-0.1413	msa_change_income
0.2587	OH	0.0909	msa_unemployment_rate
-0.2830	TX	0.0470	msa_logpop
-0.5574	One-to-four family dwelling (other than manufactured housing)	0.0639	st_hownership
0.2481	Preapproval was requested	-0.0005	st_construction
0.2169	Owner-occupied as a principal dwelling	-0.1606	appl_to_msa_income
0.2036	Atlanta, Sandy Springs, Roswell - GA	0.0097	mortgage_healthscore
1.1090	Burlington - NC	-0.0154	vantage_score_2014
0.2690	Charlotte, Concord, Gastonia - NC, SC	0.0000	average_dept
-0.6403	Home purchase		

Of the more than 500 variables, there are less than 50 variables that are statistically meaningful for the predicting applicant to be black. The numerical parameters given in the HMDA basically tend to survive, although their absolute values are very small. Many other local variables have been removed.This means that blacks are distributed throughout the country.Variables about co-applicants seem to be able to predict whether the applicant is black or not because the gender and race variables of co-applicant survived and their coefficient is estimated to be high. These estimates suggest that co-applicants of black applicants tend to be either black or not co-applicants. If the preapproval is requested and the applicant is living in the house as a host, the probability of being black is increased, while the probability becomes low when the purpose of the loan is a home purchase, the property type is not a manufactured housing, and the applicant has a mortgage. Whether it is a HOEPA loan seems to have nothing to do with it. In addition, ten of the 17 variables added by the researchers survived. In particular, with respect to the variables that characterize the area in which applicants live, (appl _to _msa income is the average income of the applicant's area relative to the applicant's income), the higher the income relative to other races in the applicant's area, Also, the higher the vantage score, the higher the local development investment, the higher the income, the less likely the person is black. On the other hand, applicants who live in areas where the population of their place of residence is high or where the unemployment rate has increased is more likely to be black.

4.1.2. Asian

The estimated model $\widehat{m}_A(Z)$, which predicts whether applicant is asian or not by applying LASSO penalty to logistic regression, is as follows.

Table 5: Selected variables for the asian group from the logistic regression with LASSO penalty

coeff	Variables	coeff	Variables
0.0066	tract_to_msamd_income	0.9430	Worcester - MA, CT
0.0001	population	0.2609	Home purchase
0.0236	minority_population	-0.2920	Refinancing
0.0002	number_of_owner_occupied_units	0.0833	Not applicable.2
-0.0005	number_of_1_to_4_family_units	0.7433	co_app_gender_Male
0.0014	loan_amount_000s	0.8709	co_app_gender_No co-applicant
0.0000	hud_median_family_income	5.3838	co_app_race_Asian
-0.0066	applicant_income_000s	0.4643	co_app_race_No co-applicant
-0.0466	IN	0.7534	co_app_ethn_No co-applicant
-0.6704	MD	-1.5209	applicant_gender_Information not provided by applicant
-0.0583	NY	-0.0040	applicant_gender_Male
0.0855	OR	-2.4329	applicant_ethn_Hispanic or Latino
0.8549	PA	-0.9874	applicant_ethn_Information not provided by applicant
0.1173	WA	0.0000	msa_average_income
0.0544	One-to-four family dwelling (other than manufactured housing)	0.0168	msa_change_income
-0.5395	Owner-occupied as a principal dwelling	-0.0385	msa_unemployment_rate
-0.4303	Austin, Round Rock - TX	0.0342	msa_logpop
-0.4897	Boston - MA	-0.0695	st_hownership
0.3378	Charlotte, Concord, Gastonia - NC, SC	-0.0002	st_construction
0.6124	Dallas, Plano, Irving - TX	0.0014	loan_to_income
-0.2845	Denver, Aurora, Lakewood - CO	0.1994	appl_to_msa_income
-0.1825	San Diego, Carlsbad - CA	0.0031	mortgage_healthscore
0.2381	Seattle, Bellevue, Everett - WA	-0.0021	vantage_score_2014
0.7109	Stockton, Lodi - CA	-0.0001	average_dept

-0.3498	Washington, Arlington, Alexandria - DC, VA, MD, WV		

There are about 50 variables that survive when LASSO regression is performed with the variable indicating whether the applicant is asian or not as a dependent variable. The selected variables are not exactly the same as when the policy variable is black, and there are slight variations. For Asian people, the numerical values given in the HMDA basically tend to survive, although the absolute values themselves are very small. Many other local variables have been removed. This means that Asians are distributed all over the United States.Asian also shows no co-applicant or co-applicant likewise asian. Asians have no correlation with whether they requested preapproval or not. As opposed to black people, if the applicant lives in the house as a host, the probability of being Asian decreases and the probability of being Asian is increased if the property type is not manufactured hosing. The information about lien status was categorized as a meaningless variable in predicting whether the applicant is asian or not. Instead, it is shown that the lending purpose variable is statistically significant for the prediction.If the purpose of the loan is a home purchase, the probability of being Asian is increased, and if it is refinancing, the probability of being Asian is low. Asian also do not have any relationship with the HOEPA loan status variable. In addition, 11 of the 17 variables added by the researchers survived. In particular, with regard to variables representing the characteristics of the area in which applicants live, the higher the unemployment rate, the higher the vantage score of the state they are in, the higher the investment in development, the higher the home ownership rate in the region, Chances of being asian are low. On the other hand, the higher the population and the higher the mortgage health score of the area, the more likely it is to be Asian. Personally, the higher the loan amount is compared to the income, the higher the relative income in the locality, the greater the probability of being Asian.

4.1.3 Others

The estimated model $\widehat{m}_O(Z)$, which predicts whether applicants belong to others group or not by applying LASSO penalty to logistic regression, is as follows.

Table 6: Selected variables for the others group from the logistic regression with LASSO penalty

coeff	**Variables**	**coeff**	**Variables**
0.0028	tract_to_msamd_income	-0.2468	co_app_gender_Information not provided by applicant
0.0000	population	0.1063	co_app_gender_No co-applicant
0.0100	minority_population	4.2191	co_app_race_American Indian or Alaska Native
0.0002	number_of_owner_occupied_units	4.7188	co_app_race_Information not provided by applicant

0.0000	number_of_1_to_4_family_units	3.8021	co_app_race_Native Hawaiian or Other Pacific Islander
-0.0005	loan_amount_000s	0.3929	co_app_race_No co-applicant
0.0000	hud_median_family_income	-0.2879	co_app_ethn_Hispanic or Latino
-0.0015	applicant_income_000s	-1.3508	co_app_ethn_Information not provided by applicant
0.3466	AL	0.5630	co_app_ethn_No co-applicant
-0.2368	CA	2.7293	applicant_gender_Information not provided by applicant
-0.0135	FL	-0.1113	applicant_gender_Male
0.5116	GA	1.4262	applicant_ethn_Hispanic or Latino
-0.0372	LA	4.6346	applicant_ethn_Information not provided by applicant
-0.1695	NC	0.0000	msa_average_income
0.5037	One-to-four family dwelling (other than manufactured housing)	-0.0349	msa_unemployment_rate
0.0055	Owner-occupied as a principal dwelling	-0.1207	msa_logpop
0.4876	Austin, Round Rock - TX	-0.0358	st_hownership
-0.1292	Miami, Miami Beach, Kendall - FL	0.0000	st_construction
-0.0520	San Jose, Sunnyvale, Santa Clara - CA	0.0043	loan_to_income
0.4842	Refinancing	0.1018	appl_to_msa_income
-0.0780	Secured by a first lien	-0.0021	mortgage_healthscore
0.3064	Secured by a subordinate lien	-0.0024	vantage_score_2014
		0.0000	average_dept

Others include all cases except white, black, and Asian, including instances of Native American, Alaska Native, Native Hawaiian or other Pacific Islander, Not applicable, no information is provided. The total number of variables selected through the LASSO penalty is less than 50. In the case of Others, too, the numerical values given in HMDA basically tend to survive, although the absolute value itself is very small. Many other local variables have been removed. This means that applicants belonging to Others are distributed throughout the United States. First, when there is no co-applicant and when the co-applicant belongs to the others group as well, the applicant is more likely to belong to the others group. In the case of applicants belonging to Others, there is no relation as to whether the preapproval was requested and the house owner are living in the house. If the property type is not manufactured hosing, the probability of belonging to the Others group increases. If an applicant's mortgage is set up by the first lien, the applicant is less likely to belong others. If the mortgage is set by the subordinate lien and the purpose of the loan is refinancing, the applicant's probability of belonging to others increases.In the Others group, it seems that there is absolutely no relation to whether it is a HOEPA loan. In addition, ten of the 17 variables added by the researchers survived. In particular, with regard to the variables that characterize the area in which applicnats live, the higher the unemployment rate, the higher the vantage score or the mortgage health score of their state, the higher the home ownership rate of the area, the more the person is, the less likely it is for the applicant to belong to the others. On the other hand, the higher the loan amount and the higher the relative income in the area, the higher the probability of belonging to the others.

4.1.4 Hispanic

The estimated model $\widehat{m}_H(Z)$, which predicts whether applicants are hispanic or not by applying LASSO penalty to logistic regression is as follows.

Table 7: Selected variables for the hispanic group from the logistic regression with LASSO penalty

coeff	**Variables**	**coeff**	**Variables**
0.0017	tract_to_msamd_income	2.0586	Jefferson City - MO
0.0000	population	-1.8726	Kennewick, Richland - WA
0.0324	minority_population	-0.5308	Killeen, Temple - TX
-0.0001	number_of_owner_occupied_units	-0.7047	Knoxville - TN
0.0000	number_of_1_to_4_family_units	0.0606	Lake Havasu City, Kingman - AZ
0.0008	loan_amount_000s	-0.4771	Lansing, East Lansing - MI
0.0000	hud_median_family_income	0.0042	Las Vegas, Henderson, Paradise - NV
-0.0072	applicant_income_000s	0.6272	Lincoln - NE
-0.7445	AL	-1.5980	Longview - TX
0.3439	AR	-0.6401	Los Angeles, Long Beach, Glendale - CA
0.4584	AZ	0.1016	Lubbock - TX
0.1931	CA	0.1156	Madera - CA
1.1108	CO	0.6868	Madison - WI
0.0386	FL	1.0471	McAllen, Edinburg, Mission - TX
-1.2215	GA	0.4167	Memphis - TN, MS, AR
1.3677	ID	-0.5868	Merced - CA
-0.1839	IL	1.1309	Miami, Miami Beach, Kendall - FL
-0.3923	IN	1.1653	Milwaukee, Waukesha, West Allis - WI
0.3303	KS	-0.0516	Minneapolis, St. Paul, Bloomington - MN, WI
-1.5774	KY	0.4580	Myrtle Beach, Conway, North Myrtle Beach - SC, NC
-0.6697	LA	0.8330	Napa - CA
-0.0242	MA	-0.3471	Naples, Immokalee, Marco Island - FL
-1.2558	MD	0.8625	New Haven, Milford - CT
-0.1190	ME	0.7993	North Port, Sarasota, Bradenton - FL
-1.2943	MS	1.4271	Ocala - FL
-0.4360	NC	0.4344	Odessa - TX
0.8424	NE	0.2898	Ogden, Clearfield - UT

0.8146	NM	0.1460	Orlando, Kissimmee, Sanford - FL
0.0004	NV	2.5815	Oshkosh, Neenah - WI
-0.7300	NY	1.5347	Palm Bay, Melbourne, Titusville - FL
-0.4393	OH	1.7965	Panama City - FL
0.4892	OK	0.7926	Pensacola, Ferry Pass, Brent - FL
-0.0451	OR	0.6976	Pittsburgh - PA
0.2408	PA	0.5092	Pocatello - ID
0.0032	SC	-0.4093	Port St. Lucie - FL
-0.2833	TN	0.8051	Portland, Vancouver, Hillsboro - OR, WA
0.9458	TX	-1.4393	Providence, Warwick - RI, MA
1.4246	UT	-1.6781	Raleigh - NC
0.1852	VA	2.8379	Reading - PA
1.0336	WA	-0.6921	Reno - NV
0.0937	WI	-1.3212	Richmond - VA
-0.7095	One-to-four family dwelling (other than manufactured housing)	0.0838	Riverside, San Bernardino, Ontario - CA
0.7838	Not applicable	2.5157	Rochester - MN
0.0503	Preapproval was requested	-0.2626	Sacramento, Roseville, Arden-Arcade - CA
-0.1181	Owner-occupied as a principal dwelling	0.3773	Salt Lake City - UT
1.2599	Akron - OH	0.2792	San Antonio, New Braunfels - TX
2.7633	Albany - GA	-0.6644	San Jose, Sunnyvale, Santa Clara - CA
1.6979	Allentown, Bethlehem, Easton - PA, NJ	0.7726	San Luis Obispo, Paso Robles, Arroyo Grande - CA
2.7917	Ann Arbor - MI	0.2613	Santa Cruz, Watsonville - CA
0.5756	Atlanta, Sandy Springs, Roswell - GA	0.4005	Santa Rosa - CA
-0.3090	Austin, Round Rock - TX	1.7383	Savannah - GA
0.2436	Bakersfield - CA	1.7220	Scranton, Wilkes-Barre, Hazleton - PA
0.5994	Baton Rouge - LA	3.1480	Sioux City - IA, NE, SD
0.7048	Beaumont, Port Arthur - TX	-0.0229	St. George - UT
0.7336	Bellingham - WA	-0.1319	St. Louis - MO, IL
-0.2321	Birmingham, Hoover - AL	-0.1309	Stockton, Lodi - CA
0.4390	Bloomsburg, Berwick - PA	0.4548	Syracuse - NY
-0.0870	Boise City - ID	-1.5338	Tucson - AZ
2.4679	Bridgeport, Stamford, Norwalk - CT	0.1709	Tulsa - OK
0.2412	Brownsville, Harlingen - TX	0.3507	Tyler - TX
-0.0883	Buffalo, Cheektowaga, Niagara Falls - NY	-1.7897	Vallejo, Fairfield - CA
2.8512	Burlington - NC	0.1499	Virginia Beach, Norfolk, Newport News - VA, NC
0.9634	Cape Coral, Fort Myers - FL	2.2456	Waco - TX

1.0335	Carson City - NV	0.8830	Washington, Arlington, Alexandria - DC, VA, MD, WV
0.8956	Chicago, Naperville, Arlington Heights - IL	2.0923	Waterloo, Cedar Falls - IA
0.7020	Chico - CA	-0.9554	Wichita Falls - TX
-0.4288	Cincinnati - OH, KY, IN	1.0254	Wilmington - NC
2.7407	Cleveland - TN	0.5170	Worcester - MA, CT
0.6960	Cleveland, Elyria - OH	2.1675	Yuma - AZ
1.6472	College Station, Bryan - TX	0.4711	Home purchase
-2.8204	Colorado Springs - CO	0.2617	Refinancing
-0.6728	Columbia - SC	0.1260	Not applicable.2
-0.6761	Columbus - OH	-1.0002	Secured by a first lien
-0.0280	Corpus Christi - TX	-0.5139	Secured by a subordinate lien
-0.2768	Dallas, Plano, Irving - TX	-0.0116	Not a HOEPA loan
0.6541	Davenport, Moline, Rock Island - IA, IL	1.1947	co_app_gender_Information not provided by applicant
0.7620	Dayton - OH	0.1263	co_app_gender_Male
-0.7490	Deltona, Daytona Beach, Ormond Beach - FL	0.3597	co_app_gender_No co-applicant
-0.3906	Denver, Aurora, Lakewood - CO	-0.0651	co_app_race_American Indian or Alaska Native
-0.0940	Des Moines, West Des Moines - IA	0.7313	co_app_race_Asian
2.8729	Dothan - AL	-1.6903	co_app_race_Native Hawaiian or Other Pacific Islander
1.5602	Durham, Chapel Hill - NC	0.3518	co_app_race_No co-applicant
1.4360	El Paso - TX	4.1414	co_app_ethn_Hispanic or Latino
1.4200	Farmington - NM	-0.8907	co_app_ethn_Information not provided by applicant in mail, Internet, or telephone application
-0.4903	Fayetteville - NC	0.5051	co_app_ethn_No co-applicant
-1.5599	Flagstaff - AZ	-3.3733	applicant_gender_Information not provided by applicant
0.2075	Fort Smith - AR, OK	0.1155	applicant_gender_Male
-0.7380	Fresno - CA	0.9564	applicant_race_American Indian or Alaska Native
-0.1520	Gainesville - GA	-2.8996	applicant_race_Asian
0.2815	Grand Junction - CO	-3.7327	applicant_race_Black or African American
1.4375	Grand Rapids, Wyoming - MI	0.1371	applicant_race_Information not provided by applicant
0.9713	Greeley - CO	0.5454	applicant_race_Native Hawaiian or Other Pacific Islander

0.6296	Greenville, Anderson, Mauldin - SC	0.0000	msa_average_income
0.3115	Hickory, Lenoir, Morganton - NC	0.0301	msa_change_income
-0.0877	Houston, The Woodlands, Sugar Land - TX	-0.0005	msa_unemployment_rate
1.2391	Huntsville - AL	0.2226	msa_logpop
1.7262	Idaho Falls - ID	-0.0759	st_hownership
0.1801	Indianapolis, Carmel, Anderson - IN	-0.0006	st_construction
3.3276	Jackson - MI	-0.0034	loan_to_income
		0.0742	appl_to_msa_income

In the case of Hispanic variables, there were many local variables surviving even when the LASSO penalty was imposed. As a result, there were about 200 independent variables. This means that Hispanic applicants live more or less in specific areas. In the case of Hispanic, too, the numerical variables given in the HMDA basically tend to survive, although the absolute value itself is very small. First, when the co-applicant is absent and the co-applicant is also Hispanic, the probability that the applicant is hispanic increases. In addition, even if the co-applicant is Asian, there is an increased probability that the applicant is Hispanic. If the landlord is living in the house, if a mortgage is set up, and if he lives in a manufactured housing, the probability of the applicant being hispanic is reduced.The probability of being hispanic increases if the purpose of the loan is not a home imporvement and preapproval was requested. The case of Hispanic seems to have nothing to do with whether it is a HOEPA loan. In addition, eight of the 17 variables added by the researchers survived. In particular, with regard to the variables that characterize the region in which applicants live, the higher the development investment and home ownership rate of the state they are in, the higher the unemployment rate in their area and the higher the loan amount of the applicant's personal income, the possibility that such applicants are hispanic decreases. On the other hand, the higher the income growth rate, the larger the population, and the higher the relative income within the applicant's individual territory, the greater the probability that the applicant is Hispanic. All variables related to creditworthiness were removed.

4.2. Variable selection for the Outcome variable

Now, we include the artificial variables in a set of independent variables instead of the originally given racial and ethnicity variables, and choose variables that are important for predicting whether the mortgage loan approval was denied. In this process, about 100 variables were selected when a variable was selected via LASSO.

Table 8: selected variables for the outcome variable from the logistic regression with LASSO penalty

coeff	**Variables**	**coeff**	**Variables**
0.9581	exp_Black	0.7077	Fresno - CA
0.6572	exp_Others	0.3889	Houston, The Woodlands, Sugar Land - TX

-0.0016	tract_to_msamd_income	-0.3725	Indianapolis, Carmel, Anderson - IN
0.0000	population	0.2162	Kansas City - MO, KS
0.0031	minority_population	0.1073	Kingston - NY
-0.0002	number_of_owner_occupied_units	0.0102	Knoxville - TN
0.0002	number_of_1_to_4_family_units	0.0322	Little Rock, North Little Rock, Conway - AR
-0.0003	loan_amount_000s	-0.2224	Los Angeles, Long Beach, Glendale - CA
0.0000	hud_median_family_income	-0.1231	Miami, Miami Beach, Kendall - FL
0.0015	applicant_income_000s	-0.1193	Myrtle Beach, Conway, North Myrtle Beach - SC, NC
0.0160	AL	-0.1644	Nashville-Davidson, Murfreesboro, Franklin - TN
0.1765	AR	0.3475	New York, Jersey City, White Plains - NY, NJ
-0.4790	AZ	-0.0455	Niles, Benton Harbor - MI
-0.3169	CA	0.2457	Oklahoma City - OK
-0.0125	CO	0.0100	Pittsburgh - PA
0.2444	FL	0.1451	Providence, Warwick - RI, MA
-0.0497	GA	0.3452	Reading - PA
-0.3849	IA	-0.0831	Sacramento, Roseville, Arden-Arcade - CA
-0.1447	IL	0.3625	Salem - OR
0.1672	IN	0.4571	San Antonio, New Braunfels - TX
-0.0395	MI	0.3842	Santa Rosa - CA
0.0961	MO	0.1869	Sioux City - IA, NE, SD
-0.2598	MS	-0.3580	Tucson - AZ
0.0279	NC	0.1076	Tuscaloosa - AL
-0.0063	NE	0.2463	Urban Honolulu - HI
0.1628	NV	-0.2063	Virginia Beach, Norfolk, Newport News - VA, NC
0.0396	NY	-0.0792	Worcester - MA, CT
0.0001	OH	0.1652	Youngstown, Warren, Boardman - OH, PA
0.2021	PA	-1.2588	Home purchase
0.2416	RI	-0.2294	Refinancing
0.3972	TN	0.1393	Not applicable.2
0.1028	TX	-0.6318	Secured by a first lien
-0.1981	VA	-0.1044	co_app_gender_Information not provided by applicant
-0.0562	WA	0.2891	co_app_gender_Male
-0.2503	WI	0.0567	co_app_gender_No co-applicant

-1.7861	One-to-four family dwelling (other than manufactured housing)	0.0309	co_app_race_American Indian or Alaska Native
-0.0496	Preapproval was requested	0.2079	co_app_race_Black or African American
-0.0482	Owner-occupied as a principal dwelling	0.1811	co_app_race_No co-applicant
-0.1791	Austin, Round Rock - TX	0.3815	co_app_ethn_Hispanic or Latino
-0.1448	Beckley - WV	0.0903	co_app_ethn_No co-applicant
0.0614	Birmingham, Hoover - AL	-0.0047	applicant_gender_Information not provided by applicant
-0.4777	Boston - MA	0.0741	applicant_gender_Male
-0.1219	Canton, Massillon - OH	0.0000	msa_average_income
0.0564	Cape Coral, Fort Myers - FL	0.0068	msa_change_income
0.2373	Chicago, Naperville, Arlington Heights - IL	0.0514	msa_unemployment_rate
0.2058	Dayton - OH	0.0571	msa_logpop
-0.1406	Denver, Aurora, Lakewood - CO	0.0125	st_hownership
-0.1237	Detroit, Dearborn, Livonia - MI	0.0000	st_construction
0.3163	El Paso - TX	0.1716	loan_to_income
		-0.0851	appl_to_msa_income

In this section, I identify what variables are important for predicting about the result of the mortgage loan judgements. An analysis of the meaning of each variable and its coefficient values will be described in detail in the next section. Regarding the racial and Hispanic variables with randomness, Asian and Hispanic variables were removed by LASSO. Among racial policy variables, black and others variables have positive correlation with the loan approval rejection probability. In addition, numerical variables that are basically given in HMDA data and some local variables survive. Additionally, only the variables indicating whether the housing type is the type other than manufactured housing , whether the preapproval was requested, and whether home owners live in their own home are survived. All dummy variables which are related with the loan purpose are classified as meaningful variables. Variables about mortgage lien are also important variables. The information on co-applicants' gender, race and ethnicity also survived. Among the information on co-applicants, variables representing Asian were excluded. It is also important whether the co-applicant is Hispanic or not and their gender. Information on the HOEPA loan was categorized as meaningless for predicting whether or not the loan was approved. In addition, eight of the 17 variables added by the researchers survived. Such variables indicate the applicant's economic characteristics and features of the residential area. Among the added variables, the ones that are classified as meaningless variables are the ratio of granting permanent residence to immigrants, whether to spend more than a certain portion of income on the cost of housing, the change of income in the past 5 years, variables which indicate the credit of applicants were all removed. Therefore, in this model, the variables related to credit score are lost, and it is necessary to discuss whether these variables are really meaningless to predict dependent variables. The results may be different if the data about each applicant is applied to the model rather than the average per state.

4.3. Estimation with the finally selected variables

In the second stage, I finally selected meaningful variables for predicting loan approval. We then proceed with linear logistic regression having no LASSO penalty with these variables. However, ridge penalty or LASSO penalty is basically imposed on the logistic regression program used by the researcher, so if I do not specify any options about those penalty, ridge penalty is automatically charged. Therefore, the researcher imposed the program a LASSO penalty, but instead gave a constraint constant t of 50,000, so that there would be no practical constraint. This model is $\widehat{g}(Z)$.

Table 9: Estimated coefficients from logistic regression without the LASSO penalty

coeff	**Variables**	**coeff**	**Variables**
1.2013	exp_Black	1.3212	Fresno - CA
0.8083	exp_Others	0.5162	Houston, The Woodlands, Sugar Land - TX
-0.002	tract_to_msamd_income	-1.0356	Indianapolis, Carmel, Anderson - IN
0	population	0.4641	Kansas City - MO, KS
0.0016	minority_population	2.584	Kingston - NY
-0.0002	number_of_owner_occupied_units	0.161	Knoxville - TN
0.0001	number_of_1_to_4_family_units	0.6802	Little Rock, North Little Rock, Conway - AR
-0.0002	loan_amount_000s	-0.2253	Los Angeles, Long Beach, Glendale - CA
0	hud_median_family_income	-0.3348	Miami, Miami Beach, Kendall - FL
0.0019	applicant_income_000s	-0.8441	Myrtle Beach, Conway, North Myrtle Beach - SC, NC
-0.1902	AL	-0.5305	Nashville-Davidson, Murfreesboro, Franklin - TN
0.1897	AR	0.6226	New York, Jersey City, White Plains - NY, NJ
-0.4151	AZ	-14.7486	Niles, Benton Harbor - MI
-0.3162	CA	0.5451	Oklahoma City - OK
-0.0111	CO	0.2566	Pittsburgh - PA
0.3394	FL	0.1285	Providence, Warwick - RI, MA
-0.0709	GA	1.3569	Reading - PA
-1.1806	IA	-0.2105	Sacramento, Roseville, Arden-Arcade - CA
-0.5185	IL	1.9176	Salem - OR
0.5741	IN	0.7664	San Antonio, New Braunfels - TX
0.1152	MI	1.1588	Santa Rosa - CA

0.1905	MO	2.8399	Sioux City - IA, NE, SD
-0.5673	MS	-1.3829	Tucson - AZ
0.1324	NC	1.6269	Tuscaloosa - AL
-0.5725	NE	0.6505	Urban Honolulu - HI
0.3859	NV	-0.6089	Virginia Beach, Norfolk, Newport News - VA, NC
-0.1103	NY	-0.8884	Worcester - MA, CT
0.0481	OH	0.7872	Youngstown, Warren, Boardman - OH, PA
0.1375	PA	-1.3223	Home purchase
0.6192	RI	-0.2685	Refinancing
0.5807	TN	0.1166	Not applicable.2
0.097	TX	-0.6559	Secured by a first lien
-0.2164	VA	-0.1611	co_app_gender_Information not provided by applicant
-0.1344	WA	0.3593	co_app_gender_Male
-0.3652	WI	0.0706	co_app_gender_No co-applicant
-1.9402	One-to-four family dwelling (other than manufactured housing)	0.7207	co_app_race_American Indian or Alaska Native
-0.207	Preapproval was requested	0.2299	co_app_race_Black or African American
-0.0973	Owner-occupied as a principal dwelling	0.2729	co_app_race_No co-applicant
-0.3634	Austin, Round Rock - TX	0.4314	co_app_ethn_Hispanic or Latino
-15.6396	Beckley - WV	0.0238	co_app_ethn_No co-applicant
0.581	Birmingham, Hoover - AL	-0.0957	applicant_gender_Information not provided by applicant
-0.8015	Boston - MA	0.1062	applicant_gender_Male
-14.6734	Canton, Massillon - OH	0	msa_average_income
0.4099	Cape Coral, Fort Myers - FL	0.0061	msa_change_income
0.733	Chicago, Naperville, Arlington Heights - IL	0.0554	msa_unemployment_rate
0.7543	Dayton - OH	0.0461	msa_logpop
-0.2522	Denver, Aurora, Lakewood - CO	0.0122	st_hownership
-0.4909	Detroit, Dearborn, Livonia - MI	0.0001	st_construction
0.9288	El Paso - TX	0.1813	loan_to_income
		-0.1076	appl_to_msa_income

I implemented logistic regression without LASSO penalty only with the variables that survived in the second step. As a result, the blacks or people in the others group were more likely to be denied the loan approval than white people. The majority of others groups are cases where race information is not given by the applicant. It is noteworthy that blacks are more likely to be rejected home loan applications than applicants belonging to Others.Considering the fact that many applicants belonging to Others did not provide their race information to the lender, it is more likely that the applicant will be rejected when the applicant reveals his / her race information to be black than when it does not disclose his / her race information. If so, Black applicants may find it advantageous to get a home loan approval on their own, without disclosing their race information. It is shown that there is no racial discrimination for Asian and Hispanic following the fact that such variables are classified as meaningless variables for predicting the result of mortgage loan judgment by the LASSO penalty. The surviving local variables are 25 out of about 50 states and about 30 out of about 410 MSA regions. This means that loan approval is better or worse in some areas than in other areas. Some of the MSA local variables are suddenly estimated to have a large absolute value of the coefficient. Such variables are Canton, Massillon-OH, Niles, and Benton Harbor-MI. These coefficients were estimated to be -0.0752 and -0.0434 respectively in the previous step, but when the LASSO penalty was removed, it was estimated to be -14.6734 and -14.7486, respectively. There were no big changes in other variables. The numerical parameters given in the HMDA basically showed interesting results. The coefficients of the polulation and the hum median family income variables were 0, but they did pass the LASSO penalty. Other variables tend to survive in general, but even if the LASSO penalty is removed, the estimated absolute value is very small. Although LASSO does not classify them as meaningless variables, it is necessary to discuss what kind of opinion should be given about these variables. Racial discrimination has also been applied to co-applicants. If the co-applicant is Native American, Black, and Hispanic, the applicant's loan is more likely to be rejected. If the co-applicant and the applicant are men, they are more likely to be rejected. If an applicant requested preapproval or the landlord is living in that house or a mortgage is set up by lien or an applicant do not live in a manufactured housing or the purpose of the loan is a house purchase or refinancing, then the possibility to be rejected from mortgage loan judgment decrease. On the other hand, when a lender do not know what the purpose of a loan is, the probability of being rejected increases. Even in the case of analyzing the possibility of the loan denial rate, it does not seem to matter whether it is a HOEPA loan. In addition, eight of the 17 variables added by the researchers survived. Among the variables that indicate the economic characteristics of a region, those living in areas with high unemployment rates are more likely to be denied home loan applications. The economic information of individual applicants survived. The thing which is consistent with our common sense is that the probability of being rejected increases when the loan amount is higher than the income and the probability of rejection is lower when the income is higher than others.

4.4. Double ML model

In Section 4.1, I selected the variables that are important for predicting each policy variable, and the coefficients for each variable were estimated, so the models $m_X(Z_i)$ for the policy variable were estimated. However, it is found that the estimated values are not correct when the loan approval is analyzed by including these as an independent variable according to the method proceeded in 4.2 and 4.3. Thus, I re-processed the regression analaysis using B-$\widehat{B}$, A-$\widehat{A}$,O-$\widehat{O}$,H-$\widehat{H}$ as independent variables and exploiting Y-$\widehat{Y}$ as a dependent variable. At this time, the dependent variable is a probability value between 0 and 1, and each independent variable is also a probability value between 0 and 1. Therefore, linear regression is performed. Since sample splitting is not performed, I use the results of the general OLS estimation. Then the estimated coefficient values are as follows.

Table 10: Estimated treatment effects from Double ML method with the LASSO penalty

	coeff	**std err**	**t-statistics**	**p>|t|**
constant	0.0001	0.004	-0.006	0.995
Black	0.1025	0.024	4.195	0.000
Asian	0.0080	0.023	0.354	0.723
Others	0.0870	0.027	3.242	0.001
Hispanic	0.0374	0.021	1.821	0.069

These results show that applicants in black and others are more likely to be denied approval. In the case of Hispanic, it is classified as meaningless in 95% confidence interval, but statistically significant in 90% confidence interval.

5. Analysis Results from Random Forest

According to the method described in Section 3.2, the difference between the outcome of each race and the outcome of Hispanic and Nonhispanic was analyzed as follows.

Table 11: Estimated treatment effects from Double ML method with Random Forest

	treatment effect
Black - White	0
Asian - White	0
Others - White	0
Hispanic - Nonhispanic	0

The above results suggest that there was no racial discrimination in home loan supply.

6. Comparing the result with other results from former research

Comparing results of previous studies, Black, Sceweitzer, and Mandell (1978) conducted a relatively small number of independent variables on loan, property and applicant's economic / personal information; and they analyze that an applicant's race has no influence on Housing Loan Supply. However, the variable race was statistically insignificant in the 95 % confidence interval but statistically significant in the 90 % confidence interval. This result is not consistent with Munnell et. Al

(1996) where the coefficient of Race variable is 1.00 and the t-statistics value is 3.73, which means that this variable is statistically significant. In particular, they did consider the credit score of the mortgage loan applicants. As a result, they found that if they do not consider credit rating, black people are discriminated against 18% of white people, while black people are less discriminated against white people by 8%. The analysis of the existence of racial discrimination is the same with the data analysis result of the early 2000s, Dell'Ariccia, Igan, and Laeven (2008) investigated the housing loans data by dividing them into prime mortgage market and sub-prime mortgage market; as a result, the coefficients of variables indicating whether the applicant is black were 1.526 and 1.246, respectively in each market, and significant in the 99% confidence interval.

This paper does not estimate the coefficients of each dummy variable after logistic regression analysis, but presents the results of the analysis with residual terms. Therefore, it is difficult to compare directly estimated coefficients with the results of the former studies. However, as for the presence of racial discrimination, when applying the LASSO penalty to the logistic regression and conducting variable selection, the fact that the probability that the applicant will be rejected would increase if the applicant is black or the applicant belongs to others is same as in previous studies. It seems that the reason why the results of this study are similar to those of the previous studies is that the analysis of data was done by logistic regression, which is essentially the same method with previous studies even though this study uses machine learning techniques to select useful variables. In other words, earlier studies and this paper's logistic regression analysis applying the LASSO penalty discuss the average value of the data, that is, the general tendency. Considering all the data for 2014, black applicants are generally more likely to be rejected than other applicants.

On the other hand, the results obtained by using the Random Forest method show that there is no racial discrimination at all. This may be because the Random Forest is not a model that considers all of the data in general or an average, but allows heterogeneity among the data. Unlike logistic regression analysis, Random Forest collects data with homogeneous tendencies and analyzes whether the members of the group got a loan approval or not. Thus, we analyze what is happening among mortgage loan applicants who have a particular tendency rather than a general tendency. For example, we can analyze whether there is racial discrimination among mortgage loan applicants who live in a particular area and whose income is above a certain amount, not among all applicants; and we can interpret that there is no racial discrimination within that group. This, in particular, can help to overcome the limitations pointed out in Munnell et al. (1996) – "Far more important, but discrimination occurs at the margin, not average" (20p).

So how do you know how data is categorized? Unfortunately, we cannot see what the specific model of Random Forest is, but later we can try to study the structure of the data using the Decision Tree method to investigate the data classification and racial discrimination in classified groups.

7. Conclusion

This study investigates whether there was racial discrimination in housing loan supply in 2014. For this purpose, HMDA data were obtained and other variables considered to be important for the researcher to predict the approval of the loan were added after being investigated. Then, unlike previous studies, this study first applied randomness to policy variables to estimate the treatment effect under the randomized experiment condition. Thus, this study allows the estimated treatment effect to be estimated more accurately without being affected by other external circumstances. This paper then estimates the treatment effect using two machine learning techniques, LASSO penalty and Random Forest. The analysis process and analytical results are dependent on which machine learning techniques are chosen and exploited.

When I apply the LASSO penalty to the logistic regression model, we choose independent variables that are important for predicting individual dependent variables at each stage of the research in this paper. In particular, a sufficient model to predict loan approval could be established using about 100 variables out of a total of more than 500 independent. Thus, the researcher has made it possible to consider only a manageable number of variables from a myriad of variables, which allows us to go through an efficient decision-making. More specifically, in addition to the basically given variables, information on specific residential areas, housing types, preapproval applications, loan purpose variables, lien status variables and especially co-applicants survived. It is revealed that racial discrimination has also emerged for co-applicants. Among the variables added to the data by the researchers, variables such as the relative income of the applicant in the residence area, the applicant's economic information such as the ratio of income to the loan amount, and the economic characteristics of other residential areas such as the unemployment rate in the residential area left. Interestingly, all variables related to the applicant's credibility were removed. This may be because the credit variables added to the data by the researcher are not dependent on the individual information of the applicant but are dependent on the residence area of the applicant. In this case, the housing loan supply model is largely changed from the existing model, so that certain variables are removed and the variables that were not previously considered are added to the model. This is the development of the home loan supply model. This is also very different from the existing research method which has devoted a great deal of effort to selecting variables that are considered to be necessary for predicting the dependent variables in addition to the given data. On the other hand, the intrinsic direction of exploiting the logistic regression did not change, so the results of the analysis showed that the probability of being denied from the mortgage loan judgment for blacks and applicants who belongs others group was higher as in earlier studies.

Random Forest, on the other hand, does not analyze all the data in a lump, but it classifies the data according to their individual characteristics, which is very different from logistic regression analysis. As a result, we obtained the result that racial discrimination does not exist at all. This suggests that there was no racial discrimination among applicants who have homogeneous characteristics. Unfortunately, although the Random Forest method provides a more accurate classification results, it is difficult to grasp the specific model using this method. Therefore, it is difficult for this study to account for how homogeneous data are gathered. This suggests that further studies may need to be done to analyze the structure of the data using the Decision Tree technique

References

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Black, H., R. L. Schweitzer, and L. Mandell (1978). Discrimination in mortgage lending. *The American Economic Review 68*(2), 186–191.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees.* CRC press.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, et al. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Dell'Ariccia, G., D. Igan, and L. Laeven (2008). Credit booms and lending standards: Evidence from the subprime mortgage market.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations.* CRC Press.

Haughwout, A., C. Mayer, and J. Tracy (2009). Subprime mortgage pricing: the impact of race, ethnicity, and gender on the cost of borrowing. *Brookings-Wharton Papers on Urban Affairs 2009*(1), 33–63.

Le Cun, B. B., J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems.* Citeseer.

Munnell, A. H., G. M. Tootell, L. E. Browne, and J. McEneaney (1996). Mortgage lending in boston: Interpreting hmda data. *The American Economic Review*, 25–53.