The Role of Community Banks During the Pandemic

Final Report for 2022 CSBS Data Analytics Competition

Georgetown University

Authors (In alphabetical order by last name):

Qian Leng

Yihui Liu

Zehui Wu

Faculty Advisor:

Dr. Purna Gamage

Table of Contents

1. Abstract

Community banks primarily provide services to the local business in the neighborhood communities which will enhance their personal relationship with the customers. This research project studies the influence that the community banks have on the US economy than the other regular banks and to identify the role that community banks play in recent years.

In this report, ARM and networking, hypothesis testing, regression analysis, decision trees, random forest, and boosting models were fitted to discover whether community banks will continue playing an important role in the United States. From the results of hypothesis testing, for most originating lender states, there is no obvious correlation between forgiveness amount level, current loan approval amount, and the location of business whether the headquarter is in rural or urban areas as well as different income regions, but there are still some states that don't apply this pattern. The significance of the relationship is fully based on the originating lender states. After digging deeply of the machine learning methods, it proves that though income area has a positive correlation, it is the least correlated variable while lender state and community banks are the most influential and positively correlated with the bank's forgiveness and approval amount, which could give policymakers insights about what specific aspect would let a bank be more financially powerful.

The main goal of this research project is to provide data-driven insights that would help policymakers, regulators, scholars, and others to understand the performance of community banks during the COVID-19 period. Furthermore, community banks could help small businesses maintain their labor force especially during the pandemic periods.

2. Motivations

A. In order to display the correlation between current loan approval amount as well as forgiveness amount among businesses in different income regions and in different areas, hypothesis testing methods such as two sample t-test and chi-square test are generated. The loan approval amount and forgiveness amount need to meet the business needs in a range which they can repay comfortably, thus, whether business in rural or urban areas as well as in low to moderate- or high- income regions are significant factors for the amounts.

B. Created models to predict future “approval amount” and “forgiveness amount”. It could help lender banks have a better understanding for their approval and forgiveness capacity; it may give borrowers suggestions on where to apply for a loan; It also helps policy makers have a general view of the current supply and demand for loans, then they could enact new laws to balance the market.

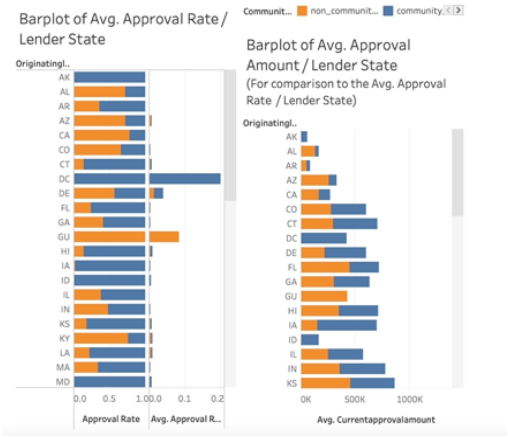
C. To discover more influence of how community banks have on the United States economic development and employment contribution during the COVID-19 period, more machine learning methods such as random forest and boosting models and interactive visualizations (link attached) are generated for policy makers and regulators.

3. Data Exploration

A. Data Cleaning:

a. Dataset I:

For this study, Paycheck Protection Program (PPP) dataset provided by CSBS has been used. The raw dataset has 19 columns and 1048575 rows. Variables “loan number”, “originating lender location id”, “cert (FDIC Bank Certificate Number)” and “date approved” are dropped firstly, since observations in the "date approved" column are almost "00:00.0" and they don’t make any sense to data analysis. Then, the rows containing missing values are dropped to improve the dataset quality. For the rest of variables, the column names are changed to make them more readable. Since the sample size of different originating states in the whole dataset is uneven, the results could be biased. For example, sample size is large for California, but small for Idaho, because of the high population in California. Therefore, three new columns which are called “employees ratio”, “approval ratio” and “forgiveness ratio” are created to avoid any inaccuracies due to biased sample sizes. The employees ratio means the number of employees of one bank as a percentage of the total number of employees in all banks. The approval ratio means the current approval amount of one bank as a percentage of the total current approval amount in all banks. The forgiveness ratio means the forgiveness amount of one bank as a percentage of the total forgiveness amount in all banks. These variables could give an intuitive explanation for the bank’s loan capacity level.



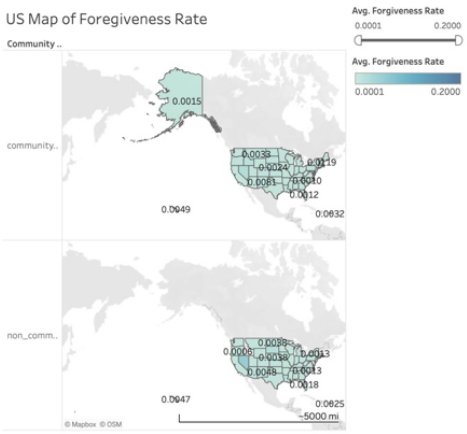
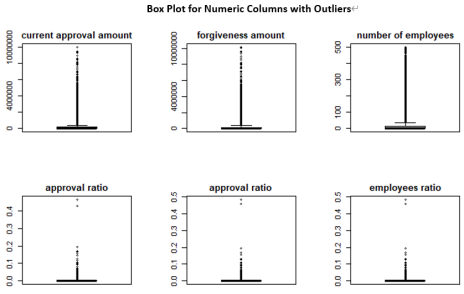


Figure 1 Figure 2

Figure 1 shows a bar plot of Avg. Loan Approval Ratio for each Lender State(left) and Avg. Approval Amount for each Lender State(right) with orange bar stands for non-community banks and blue bar for community banks. Due to the uneven sample size of different originating states, the "amount" variables could not represent the ability to give a loan for each state, but "ratio" variables could give a better explanation. Figure 2 is designed to assist readers to check by selecting a different range of Avg. Forgiveness Rate grouped by if the bank is community bank or not.

It can also be found that the original dataset has plenty of outliers *(see Figure 3)*, thus outlier processing has been done and here came out Figure 4. The final dataset is output as Dataset I, with 35955 rows and 18 columns.



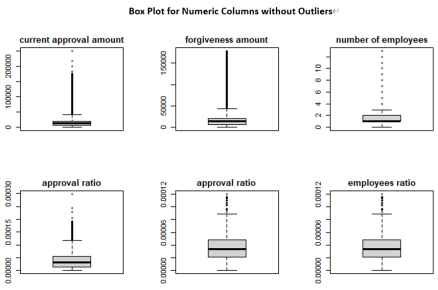


Figure 3 Figure 4

b. Dataset II:

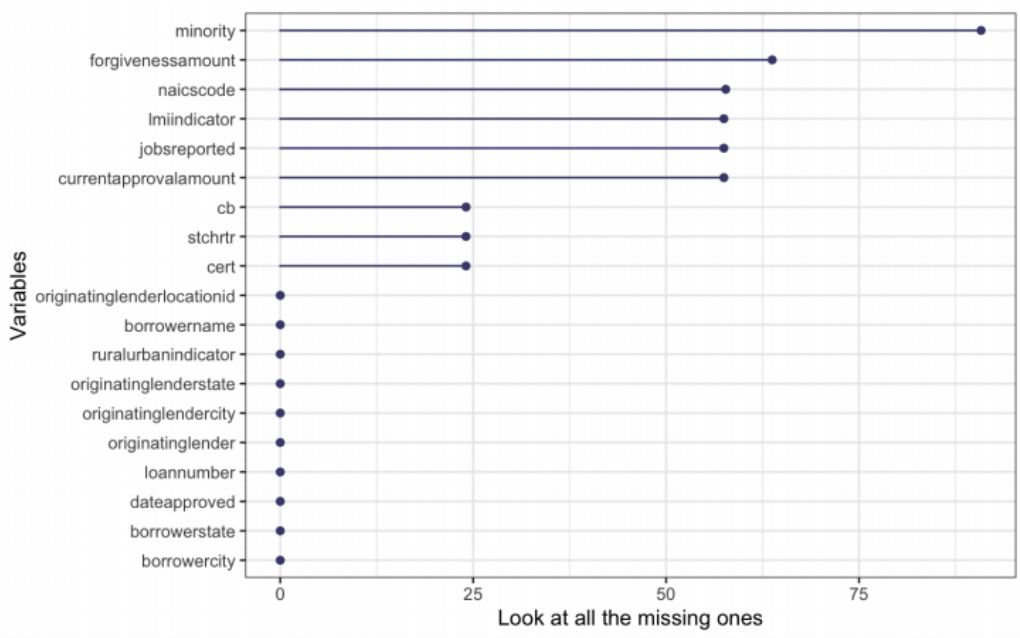


Figure 5

Figure 5 shows the summary of all the missing values of each variable in the PPP dataset. Since the "minority" variable has more than 90% missing values, which indicates the remainder of data here is not enough to use as an indicator for labeling minority owned businesses. Therefore, this column here should be dropped for the hypothesis testing analysis. After dropping the variable and repeating the cleaning steps for Dataset I, the rest of the dataset has 17 columns and 308768 rows.

B. Overview Visualizations (based on Dataset I):

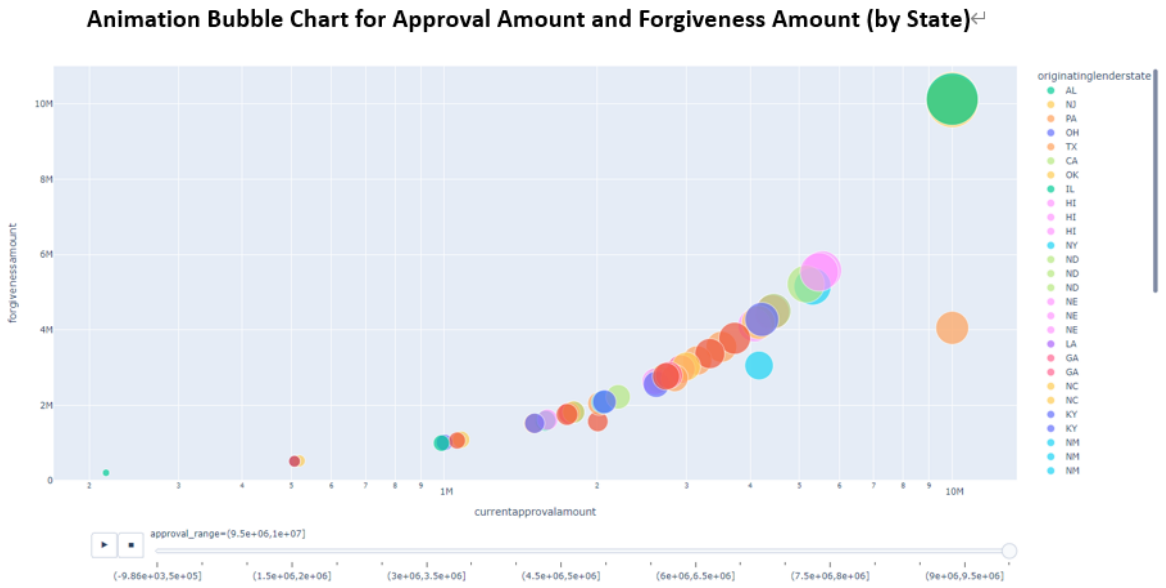


Figure 6

Figure 6 is a bubble chart made by python using plotly package. The bubble size and animation frame are decided by the current approval amount. For animation frames, the current approval amounts are divided into 20 bins, which have the same number of observations. According to the bubble chart, current approval amount and forgiveness amount have a positive relationship, and there is no obvious relationship between the originating lender state and the approval/forgiveness amounts. Therefore, it might need more specific analysis to check if there is any relationship between originating lender states and the approval/forgiveness amount.

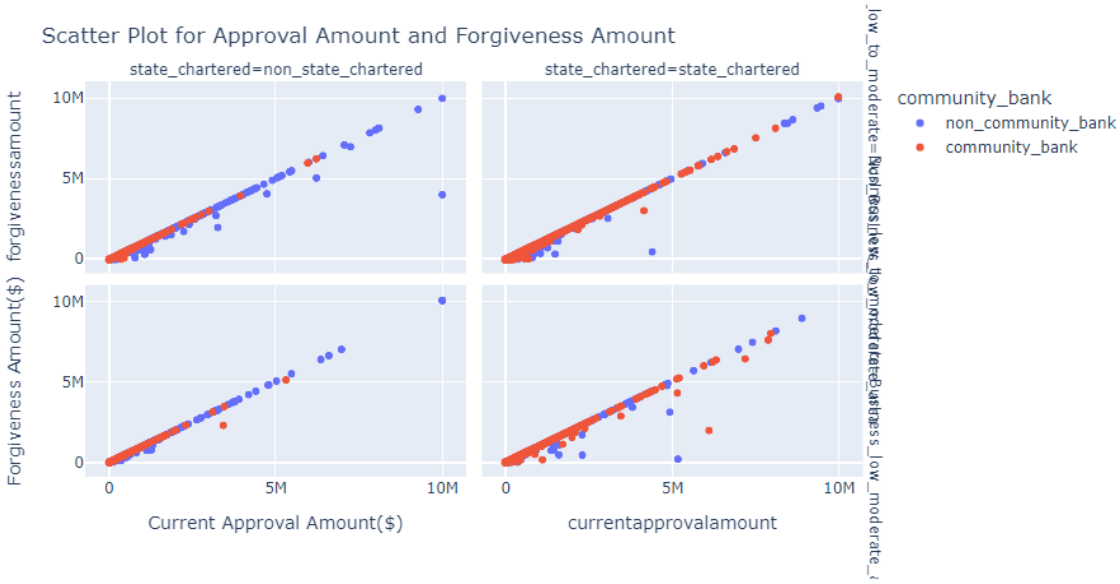


Figure 7

This scatter plot *(see Figure 7)* is mainly about current approval amount and forgiveness amount, which is divided by whether the bank is a community bank or non-community bank, whether the bank is a state chartered bank or not, and different income regions of business the bank is located in. It can be concluded that, first of all, current approval amount and forgiveness amount have a positive linear relationship. Secondly, community banks tend to have a smaller approval amount and smaller forgiveness amount compared with non-community banks. Thirdly, state chartered banks are more likely to be non-community banks, and non-state chartered banks are more likely to be community banks. Lastly, for banks in the low to moderate income area, they might have a lower approval amount and forgiveness amount.

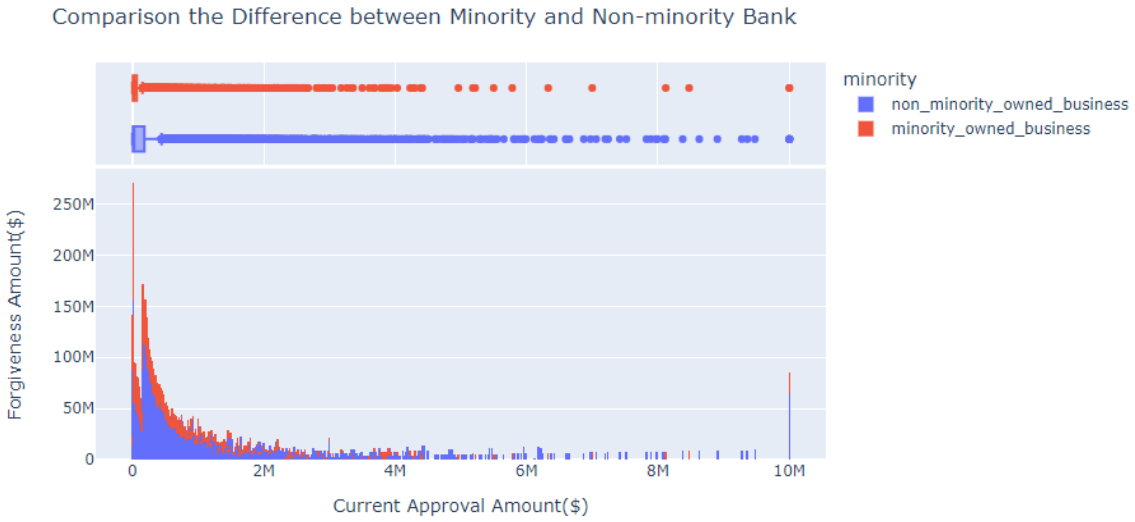


Figure 8

Figure 8 is a distribution plot that includes information about the relationship between whether the bank is minority owned and its current approval amount or forgiveness amount. The plot shows the distribution by box plot and bar chart. It could be found that no matter the banks are minority owned or non-minority owned, their approval amounts are concentrated under 2 million dollars. And since minority owned banks tend to have lower current approval amounts, thus the sum of forgiveness amounts are relatively higher when approval amounts are relatively lower. And when current approval amounts are higher than 4 million dollars, most of the banks are non minority owned. Therefore, it could be concluded that non-minority banks tend to have a better ability to give a loan.

Figure 9

Figure 9 contains two word clouds for originating lender and originating lender city. The size of the word is ordered by the word frequency. It suggests that at least in Dataset I, Prestamos CDFI is the most common or popular lender, followed by Customers Bank, U.S. Bank, Benworth Capital, Celtic Bank Corporation and BSD Capital. And for originating lender cities, the top several busiest cities are Teaneck, Phoenixville, Lake Mary, Miami, New York City and Los Angeles. Although the lists of lender banks and lender cities might be biased, at least one can believe that these banks and cities have a relatively better lending ability.

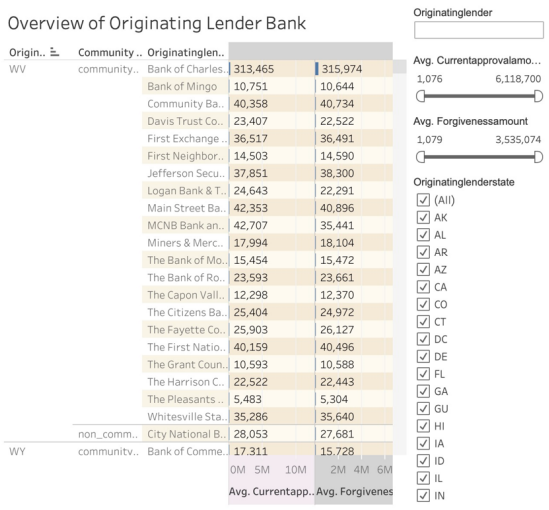


Figure 10

Figure 10 is a Tableau plot combining information of lender states, community banks, and lender banks' name with corresponding approval and forgiveness amount in total. The specific lender banks and lender states could be checked by the dropdown column on the right side. It shows clearly that for originating lender states, the number of community banks is higher than the number of non-community banks. Therefore, the total amount of both approval and forgiveness are higher in lender states. Borrowers can be offered insights that going to community banks is more likely to gain a high forgiveness and approval amount. In addition, by selecting the specific name of lender banks, the plot can offer borrowers insight about if his or her targeted lender bank is the most beneficial based on his requirements. For example, if a borrower from Wyoming is in an emergency and really needs to borrow a large amount of money over 10 thousands of dollars. It would be better for him to avoid Summit National Bank, Cowboy State Bank, and Wyoming Bank & Trust. Also, banks' regulators can check the difference between approval amount and forgiveness amount to decide to offer punishment or help.

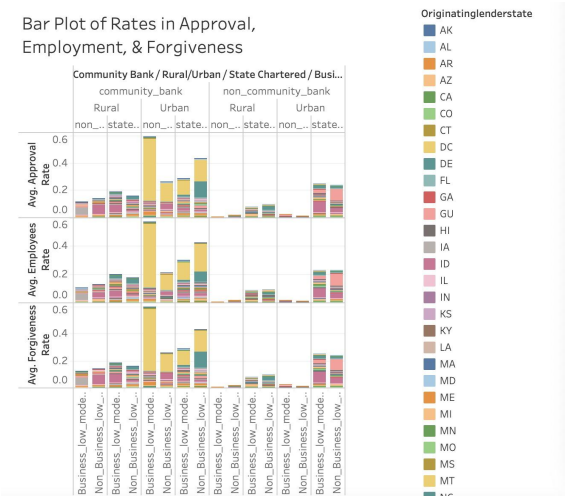


Figure 11

Figure 11 provides some combined information of community banks based on if the bank is in a rural or low income area, and if they are state chartered. Each state is marked by a distinct color as the legend shows. It is generated to give an overview of how these different categories are connected together and to understand the relationship between the forgiveness/employment/approval ratio. At the point of regulators and scholars, this plot provides information that besides the overall performance of community banks is much better than non-community banks, it also reveals that community banks in urban area with non-state chartered in low to moderate income area have the highest rates in total (e.g. Washington DC). However, for non-community banks, those in urban areas with state chartered banks have higher rates (e.g. Guam & Nevada). This can give regulators insight that “state chartered” can have a very positive influence on banks especially for non-community banks. Some insights can be given for policy makers, since the average ratio of forgiveness, employment, and approval rates are very similar in each state, which means that these three variables might have some potential relationships with each other.

4. Structure

Step 1, in order to get a better understanding of the dataset and contribute to the later analysis, ARM (Association Rule Mining) and networking analysis were generated for Dataset I. **Step 2**, based on the findings from ARM and networking analysis, three hypothesis testing were created to further explore the relationships between the variables. **Step 3**, regression analysis was done as a quantitative analysis about the target variables (current approval amount, approval ratio, forgiveness amount and forgiveness ratio). The model assessments were done here to validate the model prediction accuracy as a prediction model. **Step 4**, after learning about the limits of the regression model, decision tree models were fitted to provide more flexibility. **Step 5**, the random forest model was fitted, since it fits an ensemble of trees, which performs much better than one decision tree. **Step 6**, the boosting model was fitted to check if it might be a better choice compared to the random forest model. As a final step, conclusions were drawn from the results obtained from the analysis, and made suggestions for the policy makers.

5. Model Fitting

A. ARM & Networking:

To discover deeply about the potential relationship between variables and forgiveness, approval, and employment amount which is detected in the data exploration part, the ARM model, a rule-based machine learning method for discovering interesting relations between variables in large databases, is applied.

a. Support Rule:

$$Support = \frac{(A+B)}{Total}$$

Support rule can be explained as a measure of how frequently the collection of items ({A} and {B}) occurs together as a percentage of all transactions. Or, it can be viewed as a fraction of transactions that contain both A and B.

Table 1: Top 15 Support Rule Sets

	LHS	RHS	support	confidence	coverage	lift	count
131	{approval_rate:(0,0.1]}	{forgiveness_rate:(0,0.1]}	0.99981046523495	1	0.99981046523495	1.00017498578241	68576
132	{forgiveness_rate:(0,0.1]}	{approval_rate:(0,0.1]}	0.99981046523495	0.999985417851466	0.999825044832262	1.00017498578241	68576
129	{job_rate:(0,0.1]}	{forgiveness_rate:(0,0.1]}	0.999722987651081	0.999927087130879	0.999795885637639	1.00010206015457	68570
130	{forgiveness_rate:(0,0.1]}	{job_rate:(0,0.1]}	0.999722987651081	0.999897924960264	0.999825044832262	1.00010206015457	68570
127	{job_rate:(0,0.1]}	{approval_rate:(0,0.1]}	0.99970840805377	0.999912504557054	0.999795885637639	1.00010205866577	68569
128	{approval_rate:(0,0.1]}	{job_rate:(0,0.1]}	0.99970840805377	0.999897923471769	0.99981046523495	1.00010205866577	68569
836	{approval_rate:(0,0.1],job_rate:(0,0.1]}	{forgiveness_rate:(0,0.1]}	0.99970840805377	1	0.99970840805377	1.00017498578241	68569
837	{forgiveness_rate:(0,0.1],job_rate:(0,0.1]}	{approval_rate:(0,0.1]}	0.99970840805377	0.999985416362841	0.999722987651081	1.0001749842935	68569
838	{approval_rate:(0,0.1],forgiveness_rate:(0,0.1]}	{job_rate:(0,0.1]}	0.99970840805377	0.999897923471769	0.99981046523495	1.00010205866577	68569
123	{state_chartered}	{approval_rate:(0,0.1]}	0.843823353598974	0.999844521991501	0.843954569974777	1.00003406321271	57877
124	{approval_rate:(0,0.1]}	{state_chartered}	0.843823353598974	0.843983317778815	0.99981046523495	1.00003406321271	57877
125	{state_chartered}	{forgiveness_rate:(0,0.1]}	0.843823353598974	0.999844521991501	0.843954569974777	1.00001948056746	57877
126	{forgiveness_rate:(0,0.1]}	{state_chartered}	0.843823353598974	0.843971010688715	0.999825044832262	1.00001948056746	57877
833	{approval_rate:(0,0.1],state_chartered}	{forgiveness_rate:(0,0.1]}	0.843823353598974	1	0.843823353598974	1.00017498578241	57877
834	{forgiveness_rate:(0,0.1],state_chartered}	{approval_rate:(0,0.1]}	0.843823353598974	1	0.843823353598974	1.00018957069529	57877

Table 1 shows the top 15 sets under support rule which can validate our hypothesis that there is a very strong positive relationship between approval, forgiveness, and employment ratio.

And the most frequent category they fall is (0,0.1) which indicates the overall performance from the researched data is fair and average. Figure 12 is an interactive plot that shows the overview of the top 15 rules analysis. For some of the approval rate and forgiveness rate, they have a relationship with the state chartered variable which also gives policy makers hints about the state-chartered banks usage in increasing the approval and forgiveness rate. Also, if the bank is state chartered, it would have no direct influence on the job rate.

Figure 13 is also an interactive plot generated by R which provides a more intuitive overview of how variables are connected as discussed above.

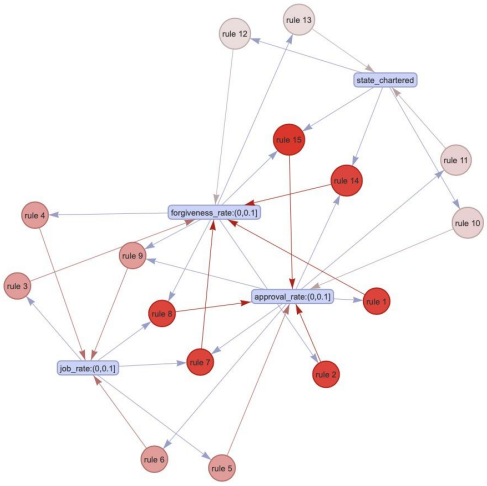


Figure 12

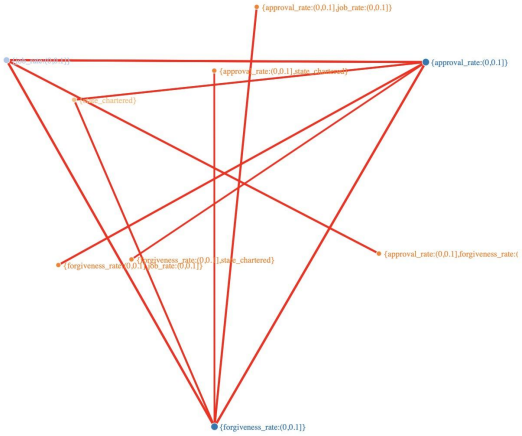


Figure 13

b. Confidence rule:

$$Confidence = \frac{(A+B)}{A}$$

Confidence rule can be explained as a ratio of the number of transactions that include all data in {B} and in {A} to the number of transactions that include all items in {A} [1]. Or, it can be viewed as how often items in B appear in transactions that contain A only.

Table 2: Top 15 Confidence Rule Sets

	LHS	RHS	support	confidence	coverage	lift	count
5	{Lender_S:WI}	{job_rate:(0,0.1]}	0.0896645234658619	1	0.0896645234658619	1.00020415603354	6150
6	{Lender_S:WI}	{approval_rate:(0,0.1]}	0.0896645234658619	1	0.0896645234658619	1.00018957069529	6150
7	{Lender_S:WI}	{forgiveness_rate:(0,0.1]}	0.0896645234658619	1	0.0896645234658619	1.00017498578241	6150
9	{Borrower_S:AL}	{job_rate:(0,0.1]}	0.0961087054775547	1	0.0961087054775547	1.00020415603354	6592
10	{Borrower_S:AL}	{approval_rate:(0,0.1]}	0.0961087054775547	1	0.0961087054775547	1.00018957069529	6592
11	{Borrower_S:AL}	{forgiveness_rate:(0,0.1]}	0.0961087054775547	1	0.0961087054775547	1.00017498578241	6592
15	{Borrower_S:WI}	{job_rate:(0,0.1]}	0.107160040239689	1	0.107160040239689	1.00020415603354	7350
16	{Borrower_S:WI}	{approval_rate:(0,0.1]}	0.107160040239689	1	0.107160040239689	1.00018957069529	7350
17	{Borrower_S:WI}	{forgiveness_rate:(0,0.1]}	0.107160040239689	1	0.107160040239689	1.00017498578241	7350
18	{Lender:Cross River Bank}	{Lender_S:NJ}	0.125749026811879	1	0.125749026811879	7.54886638784944	8625
22	{Lender:Cross River Bank}	{non_community_bank}	0.125749026811879	1	0.125749026811879	1.83309725525831	8625
25	{Lender:Cross River Bank}	{state_chartered}	0.125749026811879	1	0.125749026811879	1.18489790277442	8625
26	{Lender:Cross River Bank}	{job_rate:(0,0.1]}	0.125749026811879	1	0.125749026811879	1.00020415603354	8625
27	{Lender:Cross River Bank}	{approval_rate:(0,0.1]}	0.125749026811879	1	0.125749026811879	1.00018957069529	8625
28	{Lender:Cross River Bank}	{forgiveness_rate:(0,0.1]}	0.125749026811879	1	0.125749026811879	1.00017498578241	8625

Figure 14 shows policy makers of how often multi-layer networks appear in transactions that contain single-layer networks only. The information of which specific states of lender and borrower have the most frequent influence on each specific rate can be found. It seems that for the range of (0,0.1), the lender state of Wisconsin and borrower state of Alabama & Wisconsin have the most frequent combination together which in some level reflect these three states' economic level. Since (0,0.1) is the lowest category among 0 to 0.5 range scales. Therefore, policy makers can be suggested to establish more community banks or financial assistance to these low rate states.

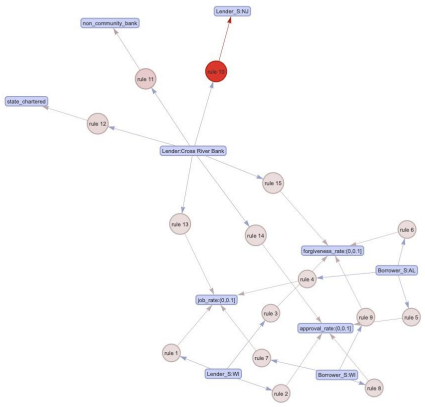


Figure 14

c. Lift Rule:

$$Lift = \frac{(\frac{(A+B)}{A})}{(\frac{B}{(Total)})}$$

Lift rule can be explained as the ratio of the confidence to expected confidence. How much our confidence has increased that B will be purchased given that A was purchased.

Table 3: Top 15 Lift Rule Sets

	LHS	RHS	support	confidence	coverage	lift	count
842	{Borrower_S:WI,community_bank,non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
2621	{Borrower_S:WI,community_bank,job_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
2626	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
2631	{Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
5107	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,job_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
5113	{Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1],job_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
5119	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
7151	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1],job_rate:(0,0.1],non_minority_owned_business}	{Lender_S:WI}	0.0806397527300296	0.935713077313483	0.0861799997084081	10.435711261765	5531
135	{Borrower_S:WI,community_bank}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
846	{Borrower_S:WI,community_bank,job_rate:(0,0.1]}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
850	{approval_rate:(0,0.1],Borrower_S:WI,community_bank}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
854	{Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1]}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
2636	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,job_rate:(0,0.1]}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
2641	{Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1],job_rate:(0,0.1]}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895
2646	{approval_rate:(0,0.1],Borrower_S:WI,community_bank,forgiveness_rate:(0,0.1]}	{Lender_S:WI}	0.0859467261514237	0.930691506157247	0.092347169371182	10.3797072708649	5895

The applied lift Rule shows the confidence level between the relationship of the increasing of multi-layer networks and the increasing of single-layer networks which means they are expected to have a positive correlation to each other. From Figure 15, it can be revealed that the response variables are all set as Lender State of Wisconsin specifically, which may be caused by the fact that the total number of community banks and non-community banks is the highest in Wisconsin among all 50 states. Therefore, focusing on the Lender State of Wisconsin, the top 15 matches show that Wisconsin has a combination of a high percentage of community banks with non-minority owned business. Since minority owned business is defined as "the business is at least 51% owned by such individuals" which reflects that in Wisconsin the survival rate for non-minority-owned business is higher than minority owned business or small business. This can give insights to policy makers in Wisconsin to help establish more small businesses to make more variety, and this is important for policy makers to expand an ownership society to all classes of the society.

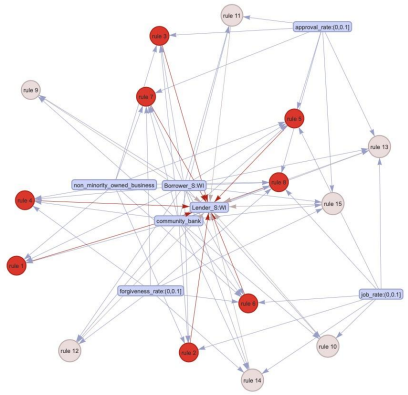


Figure 15

However, checking the table, the majority of employment ratio is still between 0 to 0.1. Since minority-owned businesses are typically smaller and have fewer employees than

non-minority-owned businesses, which can reflect the horrible survival situation of minority-owned-business in Wisconsin. Therefore, suggestions such as providing more programs on the federal, state, county and city levels which can support and foster minority business enterprises to grow and then mitigate the monopoly from large business can be given to policy makers.

B. Hypothesis Testing:

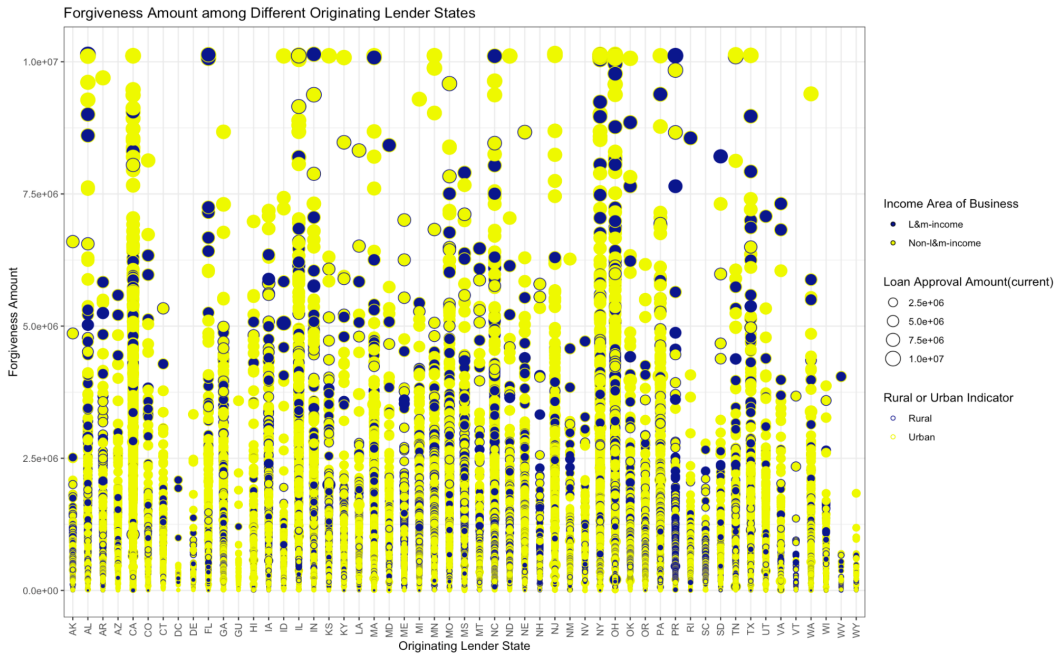


Figure 16

It can be found from Figure 16 that originating lender states with higher forgiveness amounts always have higher current loan approval amounts. The majority points in this scatterplot are marked as "urban" and "non-low to moderate-income", which means companies are more likely to locate in high income regions of urban areas. However, the dots on this scatter plot are so dense that it is difficult to analyze the relationship between each variable. Hence, more detailed charts should be generated later.

The goal is to analyze the relationship between “forgiveness amount” and “rural urban” indicator first since, from the graph above, businesses have preference on where they would like to locate their headquarters. Two sample t-test is one of the most commonly used hypothesis testing methods used for comparing the average difference between two groups, and for the

dataset analyzed here, groups are split based on the location of the business which is in a rural area or an urban area.

Before generating the two-sided t-test, a normality test should be performed to check the normality of the numerical variable. One sample Kolmogorov-Smirnov (K-S) test is a really simple and useful model for testing if a variable follows a given distribution in a population, such as a normal distribution. Thus, a K-S test is performed on the forgiveness amount using the "pnorm" parameter, which can be used to check the normality of this distribution. Null hypothesis here is that the mean forgiveness amounts of business in rural and urban areas are the same. As a result, the p-values are less than 0.05, we have enough evidence to reject the null hypothesis at 5% significance level, which indicates that the data are not normally distributed. but t-test is valid for large samples from non-normal distribution by CLT, so it is still meaningful to do a two sided t-test next.

Table 4

Two-sided t-test for forgiveness amount of business in rural and urban area	
H0: The mean forgiveness amounts of business in rural and urban area are the same.	
P-value	< 2.2e-16
95 percent confidence interval	
-77834.84	-71138.54
Mean forgiveness amount of business in rural area	Mean forgiveness amount of business in urban area
120866.6	195353.3

According to the results above, the p-value is smaller than 0.05, which means that true difference in means is not equal to 0 at 5% significance level. The forgiveness amount of business in rural and urban areas are not the same. The difference on average is about 70k

dollars, with the mean forgiveness amount of rural area and that of urban area are 120866.6 dollars and 195353.3 dollars separately.

Besides two sample t-tests, Pearson's chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables used for studying. However, the required data type of the chi-square test is categorical, thus, forgiveness amount needs to be separated into 5 groups using bin function and labeled different forgiveness amount levels as "very_large", "large", "medium", "small" and "very_small", which can be seen clearly on Figure 17 as below:

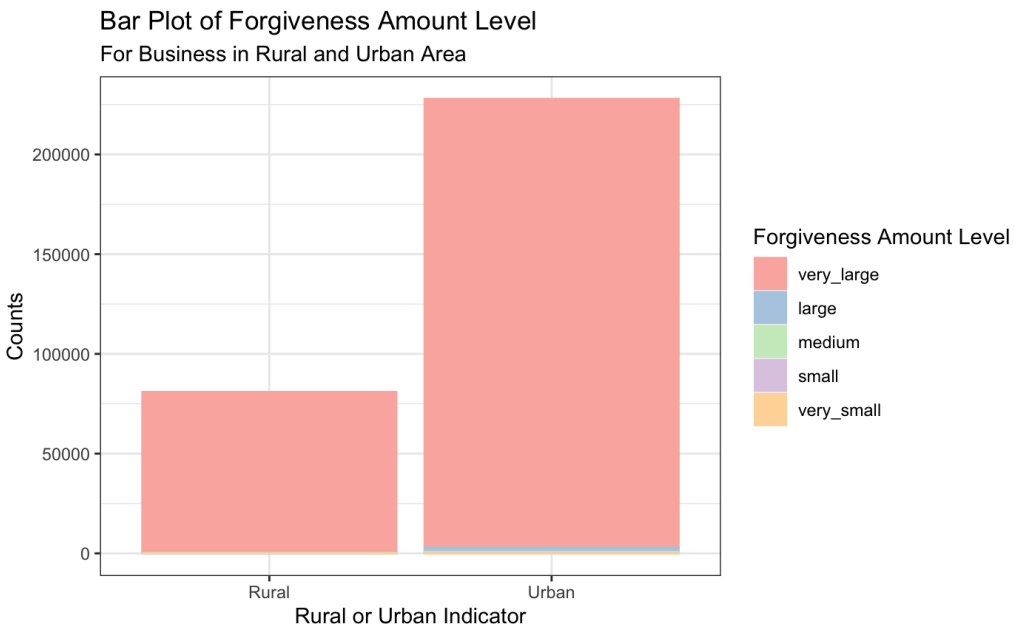


Figure 17

From Figure 17, it is clear that almost all the businesses in both rural and urban areas has a very large forgiveness level. In order to see a detailed distribution of each forgiveness amount level, two bar plots were made in Figure 18:

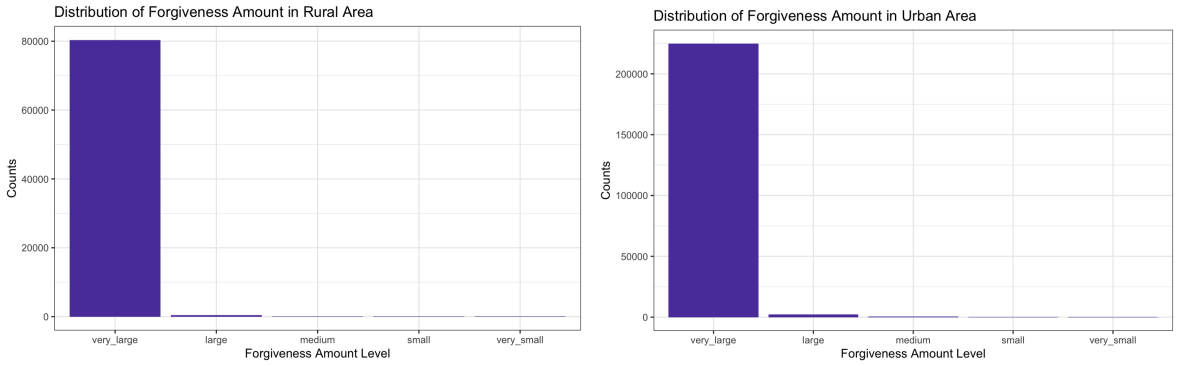


Figure 18

Bar charts here present that distribution of all forgiveness amount levels in both rural and urban areas looks similar, almost all the businesses are in high forgiveness amount level, for the rest of the business, it is more likely that they are at a large forgiveness amount level. Just very few companies are at medium or small or very small forgiveness amount level.

In order to facilitate the display the performance of chi-square test for each originating lender state, a function is created here: one can input the name of originating lender states and the output will be the p-values of the chi-square test that was done to check the forgiveness amount level in rural or urban areas or in different income regions. By checking this function, states in which businesses have high differences in loan approval amount when located in different income regions(CA, IL, NC, NJ, OH, PA) are used as an example here *(see Table 5)*.

Table 5

Originating Lender States	P-value of Chi-Square Test for forgiveness amount and different income region indicator	P-value of Chi-Square Test for forgiveness amount and rural or urban indicator
CA	0.6381022	0.007139012
IL	0.4225591	0.3667032
NC	0.3118421	0.03731832
NJ	0.03944929	0.830098
OH	0.02000392	0.1853785
PA	0.00536135	0.001197444

According to p-values in Table 5, the majority of them are larger than 0.05, which shows there is no obvious relationship between forgiveness amount level and whether the business is located in a rural or in an urban area. Similar results hold for lenders within a low to moderate- or a high-income area, but there are still some p-values of originating lender state, such as Pennsylvania, whose p-values are smaller than 0.05, which means there are significant correlation between forgiveness amount level and whether the business is located in a rural or in an urban area as well as within a low to moderate- or a high-income area. The significance of correlation is really based on the states.

The main topic for this report is the role of community banks during the pandemic, thus, "community bank" is a significant factor that needs to be taken into consideration while exploring the relationship between forgiveness amount and "rural urban" indicator. Figure 19 indicates that the PPP loan forgiveness amount of community bank in rural area is higher than non-community bank, maybe because the local community banks offer better rates and lower fees, which have led PPP lending to small businesses and support a forgiveness process that is minimally burdensome for borrowers so they can focus on preserving their businesses[2]. But this pattern doesn't apply to urban areas, it is probably because community banking organizations target small business owners and have small assets, so there aren't any community banks in urban areas. Because in rural areas, there are more than two times as many community banks as there are non-community banks; but for urban areas, the number of non-community banks is about two times as many as community banks.

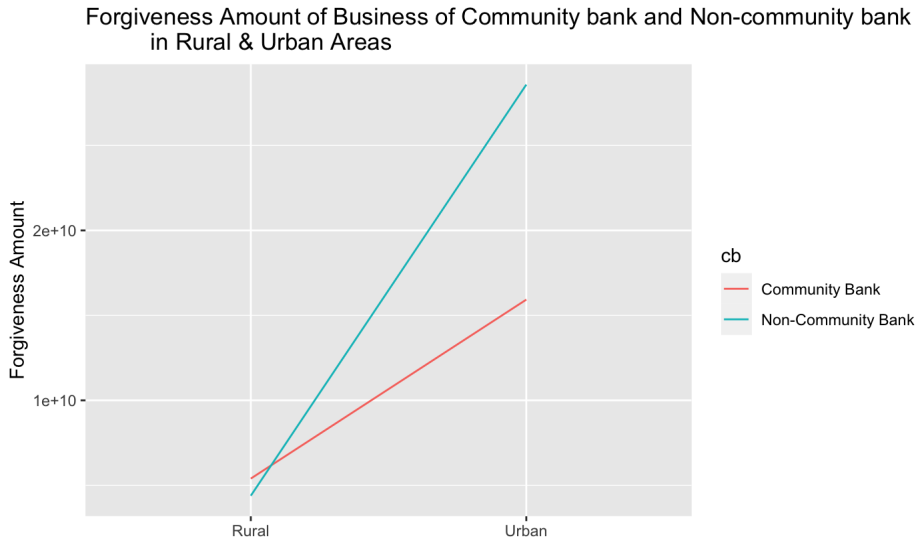


Figure 19

From Figure 16, it can be observed that businesses also have preference on high income regions rather than low or moderate income regions while choosing their location. To explore more, Figure 20 is generated to see more detailed information, that almost all the businesses in each kind of income region are at a very large forgiveness amount level.

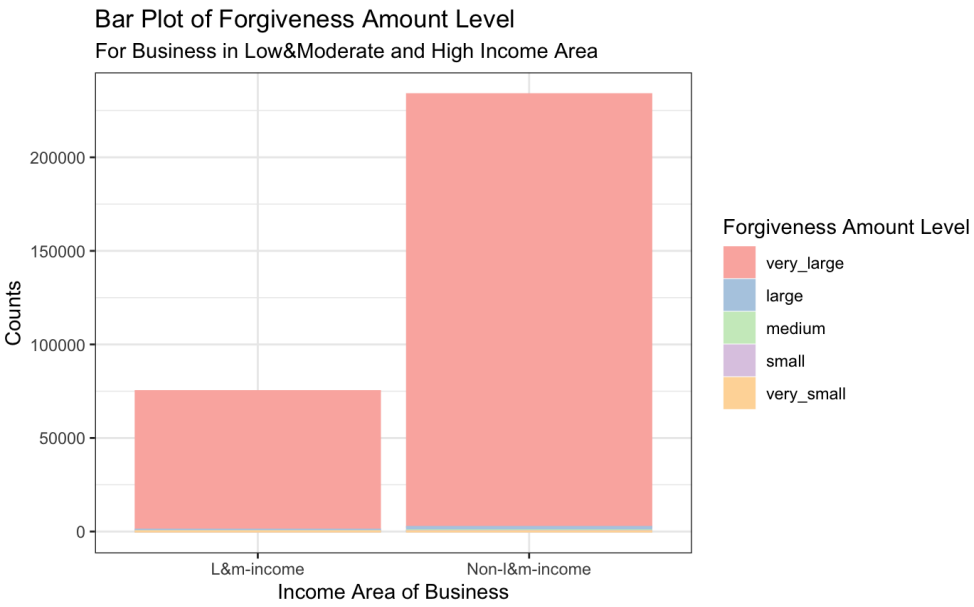


Figure 20

In addition, the current loan approval amount is an important estimator for hypothesis testing. Before generating two-sided t-test and making an exploration of the relationship between loan approval amount of business in different income regions, the normality test should be displayed first, the result is similar to the previous analysis in page 22 that although the p-values is less than 0.05, which indicates that the data are not normally distributed. T-test is valid for large samples from non-normal distribution, so it is still meaningful to do a two sided t-test next.

Table 6

Two-sided t-test for current loan approval amount of business in different income regions	
H0: The current loan approval amounts of business in low&moderate- and high-income regions are the same.	
P-value	< 2.2e-16
95 percent confidence interval	
25360.81	33733.96
Mean loan approval amount of business in low& moderate-income regions	Mean loan approval amount of business in high-income regions
198365.1	168817.7

The performance of t-test shows that p-value here is smaller than 0.05, true difference in means is not equal to 0. The current loan approval amount of businesses in low to moderate and high income regions are not the same. The difference on average is about 30k dollars, with the mean current approval amount of the low to moderate income area and that of the high income area are 198365.1 dollars and 168817.7 dollars separately.

To better understand the relationship between the loan approval amount of business in different areas as well as in different income regions of different originating lender states, Figure 21 and Figure 22 are generated.

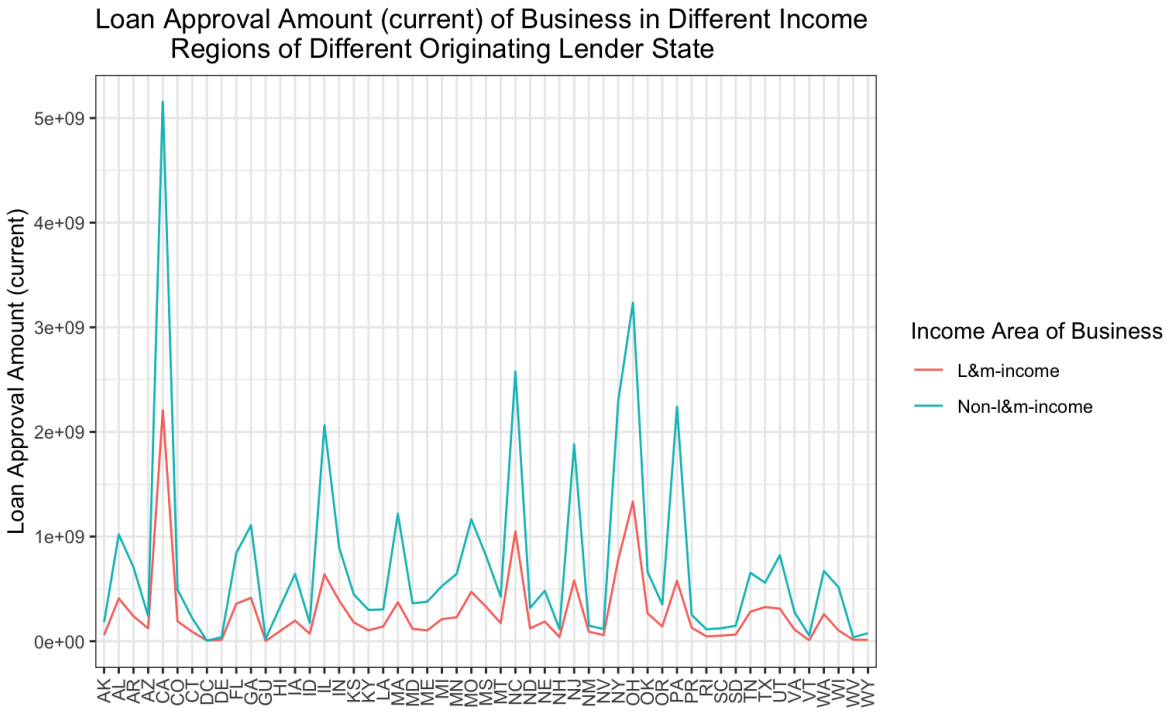


Figure 21

From Figure 21, it is clear that California and Ohio states have the highest current loan approval amount, especially for businesses not in low to moderate income areas. And it can be observed that businesses in high income area have more “current loan approval amount” than those in low and moderate income area, especially in some states, such as Ohio, California and Maryland, the “current loan approval amount” of business in high income area are two times more than those in low income area; Also, in some states, such as Washington.DC and Guam, there is not so much difference between business in high income area and that in low and moderate area.

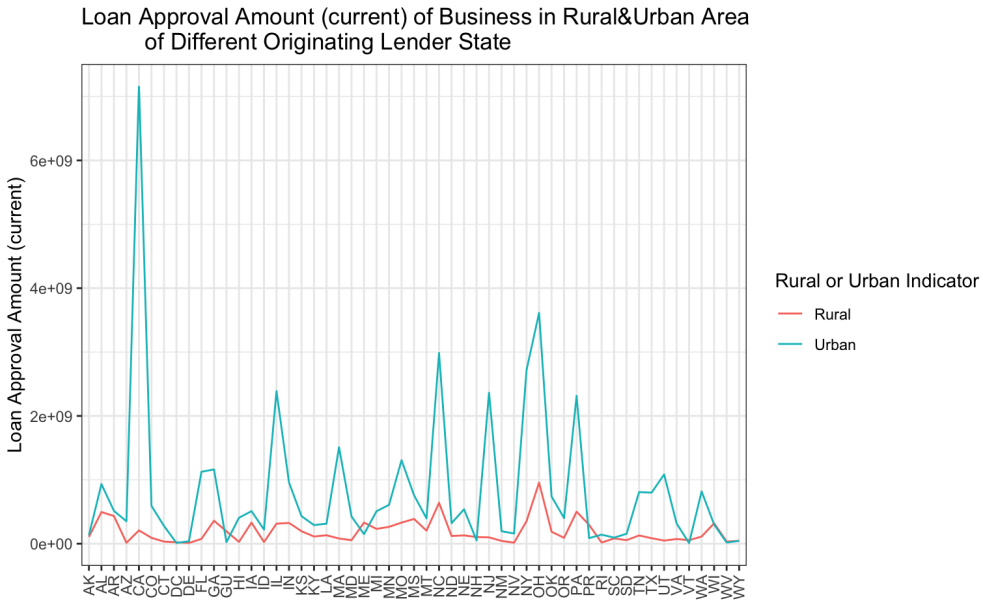


Figure 22

But according to Figure 22, it can be observed that California and Ohio state have the highest current loan approval amount, especially for businesses in urban areas. And it can be observed that businesses in urban area have much more current loan approval amount than those in rural area, especially in some states, such as California, Ohio and New York, the current loan approval amount of business in urban area are eight times more than those in rural area; Also, in some states, such as Washington.DC and Guam, there is no difference between business in urban area and that in rural area.

Then, the mean difference is calculated for each originating lender state among different income regions and different areas. TOP 6 states which have high difference and low difference are selected to make a new list and used for the following analysis. Figure 23 and Figure 24 illustrate box plots that show overall performance of the relationship between loan approval amount in different areas and different level income regions for high and low difference originating lender states.

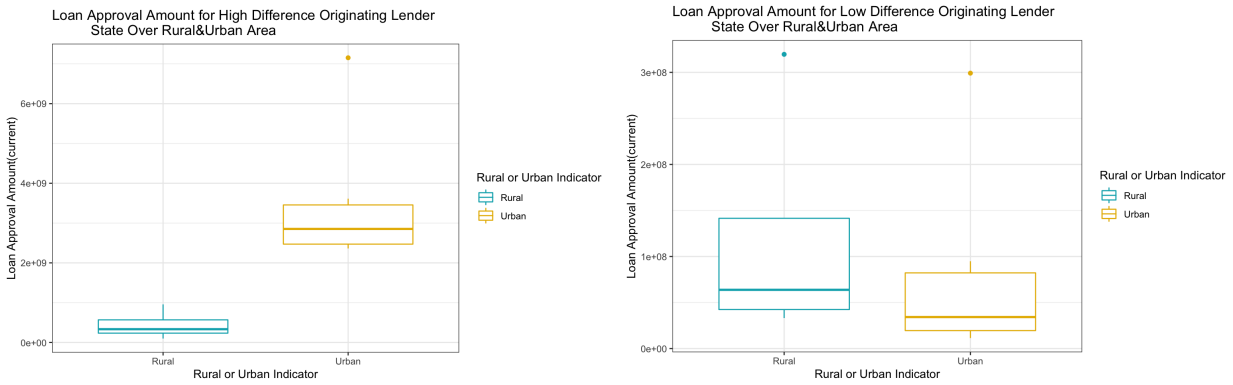


Figure 23

From Figure 23 above, the median, max, min value and outliers can be checked easily. It can be observed that in high-difference originating lender states, the overall loan approval amount in rural areas is much lower than that in urban areas; however, for low-difference states, it is surprising to see that the overall loan approval amount situation in rural areas is better than that in urban areas.

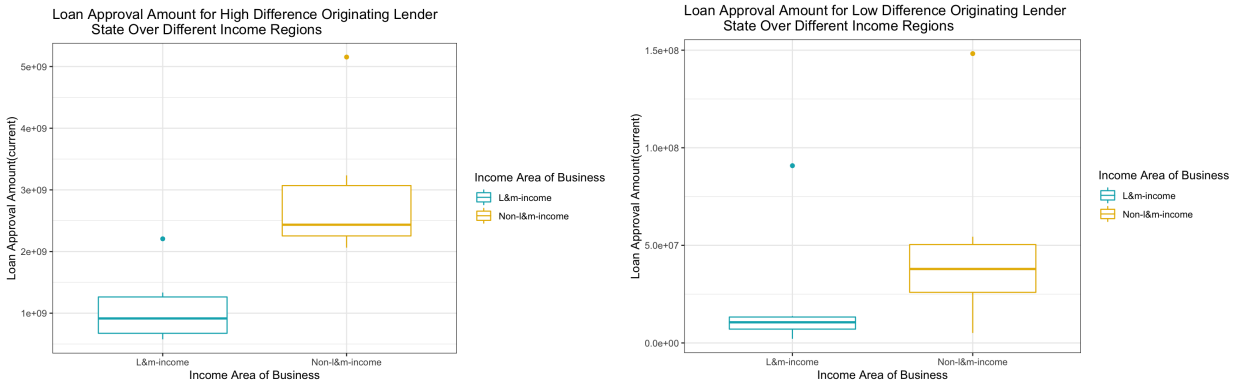


Figure 24

Figure 24 presents that the difference of overall loan approval amount situation in different income regions is larger in high-difference states than in low-difference states.

C. Regression Analysis:

Regression analysis refers to a set of statistical methods that are used to estimate the relationships between dependent and independent variables. In this project, approval amount ratio (the current approval amount of one bank as a percentage of the total current approval amount in all banks) and forgiveness amount ratio (the forgiveness amount of one bank as a percentage of the total forgiveness amount in all banks) were chosen as the dependent variables (target variables). Since the ratio variables could provide more normalized results than the amount variables, the regression model would predict the approval/forgiveness ratio instead of the approval/forgiveness amount.

Before the regression analysis, a scatter plot matrix *(Figure 25)* for all the numeric variables was generated by pairs() function in R. This could provide a general view for correlation between the variables. It suggests that the forgiveness amount has an obvious positive linear relationship with the current approval amount. And so do the approval ratio and the forgiveness ratio. It is very interesting that the number of employees in the bank does not have a significant relationship with the bank's current approval amount and forgiveness amount, which means a big bank(only valued by number of employees) is not necessary to have a good loan ability.

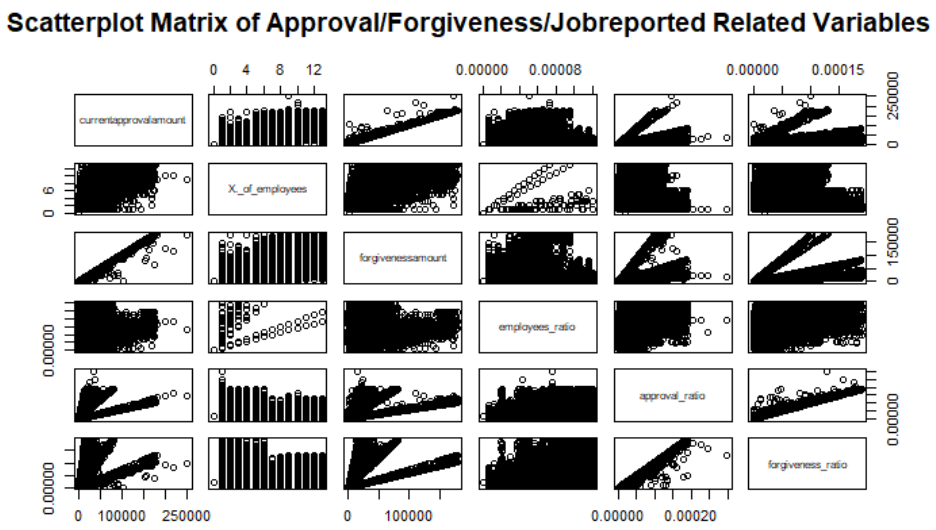


Figure 25

Since the linear regression model is the simplest and most common model used in machine learning, it is chosen as the first fitted model. Then, the Dataset I is separated into a training set(75%) and testing set(25%) randomly.

To begin with, two linear regression models were used to predict approval ratio. For the first model, the current approval ratio is chosen as the target variable, and community bank, state chartered, rural or urban, income area of business, originating lender state, employees ratio (the number of employees of one bank as a percentage of the total number of employees in all banks), forgiveness ratio and the interactive variables (employees ratio) * (forgiveness ratio) are predictor variables. For the second model, the employee ratio is changed into number of employees, forgiveness ratio into forgiveness amount. After fitting the linear regression model, Variance inflation factor (VIF) checked for the presence of multicollinearity among the predictor variables. Since the interactive variable (employees ratio) * (forgiveness ratio) and originating lender state both have a high VIF(10±1), they should be deleted from the models.

Next, the two models were fitted after removing the variables that had a high VIF. Now the new models' VIF are all under 2.5. And then the insignificant variables (95% confidence interval) were removed in each model.

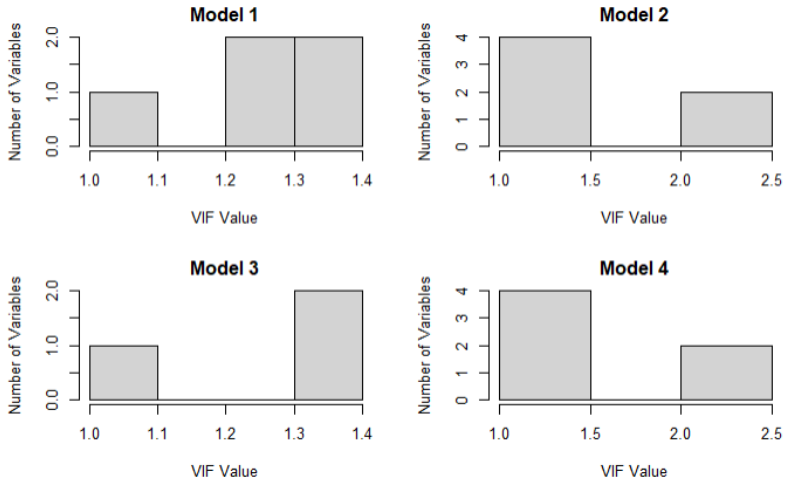


Figure 26

VIF and variables' t-test significance were checked again and then the regsubsets() function was used to find four most important variables (highlighted below) in the models. Since the prediction models for forgiveness ratio are very similar to the approval ratio prediction models, another two models could be generated. Figure 26 is a distribution plot of VIF of the four models. The models' VIF are all under 2.5, which means that the multicollinearity is very low in the models. The final four models are written below:

Model 1

approval_ratio = -0.00000002733 + 0.00000009032community_bank - 0.00000012044state_chartered + 0.00000008867urban + 0.01176577037employees_ratio + 0.98624993102forgiveness_ratio

Model 2

approval_ratio = 0.00004199523885 + 0.00000690513619community_bank - 0.00000382836596state_chartered - 0.00001300574216urban - 0.00000311085489low_to_moderate_area + 0.00000799397624number_of_employees + 0.00000000116882forgiveness_amount

Model 3

forgiveness_ratio = -0.00000018605 - 0.00000008536urban - 0.00594944601employees_ratio + 1.00678959060approval_ratio

Model 4

forgiveness_ratio = 0.0000422004596 + 0.0000068415241community_bank + 0.0000040426365state_chartered - 0.0000131922433urban - 0.0000030208423low_to_moderate_area - 0.0000082178251number_of_employees + 0.0000000011893current_approval_amount

Table 7

Summary Table for Linear Regression Model

Model	RMSE	R squared
Model 1	0.0000035	0.9919341
Model 2	0.0000338	0.2359323
Model 3	0.0000035	0.9919049
Model 4	0.0000339	0.2390237

After fitting the four models, the approval ratio and forgiveness ratio were predicted via the testing set. The RMSE(Root Mean Square Error; the standard deviation of the residuals (prediction errors))and R squared are listed in Table 7. It is obvious that Model 1 and Model 3 are the best linear regression models (have lower RMSE) to predict approval ratio and forgiveness ratio. In Model 1, banks with a high forgiveness ratio and employees ratio are more likely to have a high approval ratio. When the other predictors are fixed, 1 unit increase in the forgiveness ratio could cause the approval ratio to increase about 0.986 units. And community banks and urban banks are also prone to have a higher approval ratio, but state chartered banks

work negatively to get a higher approval ratio. In Model 3, a higher approval ratio also tends to have a higher forgiveness ratio. When the other predictors are fixed, 1 unit increase of approval ratio could cause forgiveness ratio to increase about 1 unit. And urban banks and banks with a higher employees ratio are less likely to obtain a higher forgiveness ratio.

Figure 27 and Figure 28 are model diagnostics plots for Model 1 and Model 3. For both Model 1 and Model 3, the residuals vs fitted plots reveal a linearity and homoscedasticity. It still has some outliers after outlier processing. The Q-Q plots suggest that the residuals are not normally distributed. The scale-location plots show most standardized residuals are not changing much as a function of the fitted values.

Overall, these two models could provide high-valued information to the final results.

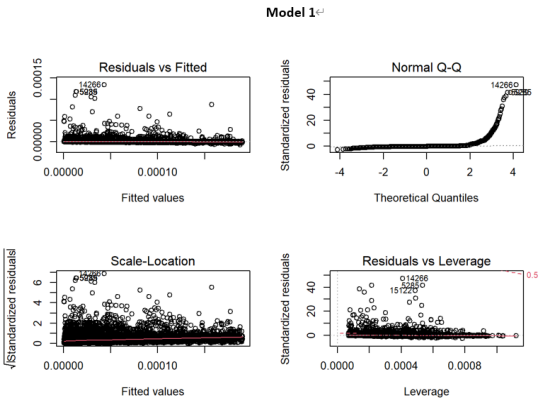


Figure 27

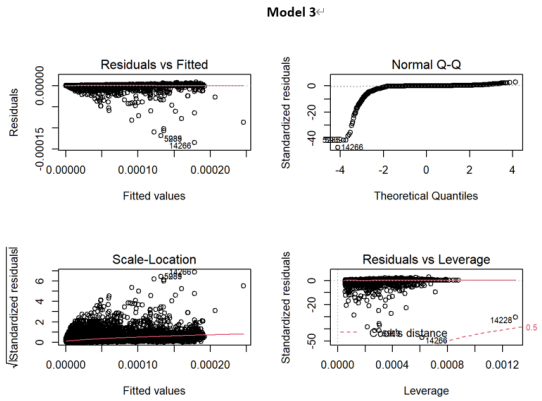


Figure 28

Even though the regression models provide more interpretability, it has a low prediction accuracy. Therefore, tree based methods which provide more flexibility and high prediction accuracy were used in the next step.

D. Decision Tree:

a. Decision Tree I:

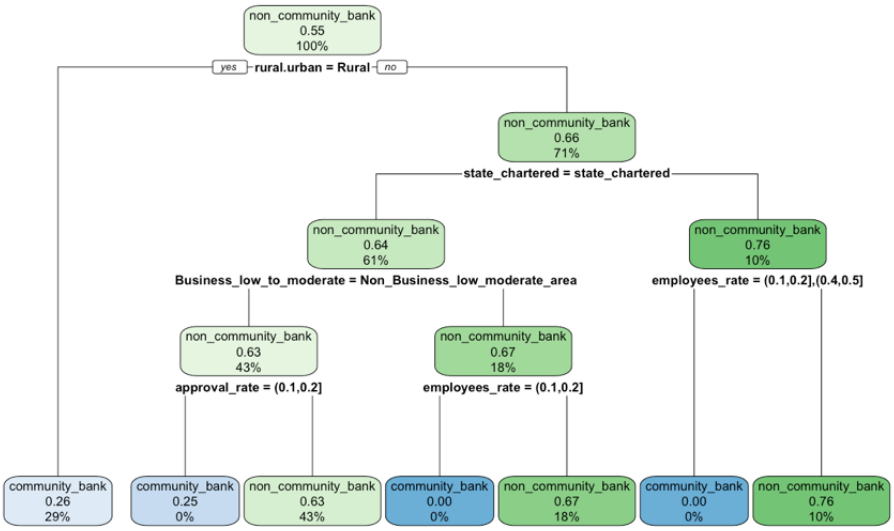


Figure 29

The above decision tree is about different ratios of forgiveness, approval, and employee. More attention is paid on non-community banks' development in this tree. Unlike community banks which are mainly located in urban areas, the decision nodes let us know that the non-community banks are mostly in rural areas where the average income is low to moderate with state-chartered. The approval rate is around 0.1 to 0.2 which is better than the majority range of 0 to 0.1 during the covid-19 period. The employment ratio performs very well (reaching 0.4 to 0.5) when the non-community bank is not chartered by states. This might be explained as some restrictions, strict regulations, or lower salary on average offered by state-chartered banks.

Therefore, the regulators of these non-community banks can revise the corresponding rules of hiring employees if they would like to expand their services.

Table 8: Prediction I on forgiveness ratio, approval ratio, and employment ratio

Forgiveness Ratio Prediction	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
Community Bank	19754	10	0	0	1
Non Community Bank	48823	0	0	0	1

Approval Ratio Prediction	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
Community Bank	19754	10	0	0	1
Non Community Bank	48822	1	0	0	1

Employment Ratio Prediction	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
Community Bank	19754	8	0	0	1
Non Community Bank	48819	1	0	0	1

Table 8 shows the prediction based on the train and test loan datasets. The distribution and overview can be shown specifically about banks and their ratio range which stands for the economic and employment level influenced by a bank for the state the bank is in. The prediction gives policymakers insights that the financial strength for community and non community banks are very similar. Compared to community banks, non-community banks have more contribution on a low level ratio while for the ratio range of 0.1 to 0.2, or a better financial contribution ratio, community banks are the major composition. It reflects that during the covid-19 period, the community banks contribute a lot to states with higher ratio of employment, approval, and forgiveness ratio, and non-community banks are beneficial more to states with lowest ratio. From this prediction, since there is a gap between 0.2 to 0.4, policy makers can make more regulations to fulfill the gap by providing more benefits and balancing the ratio distribution.

b. Decision Tree II:

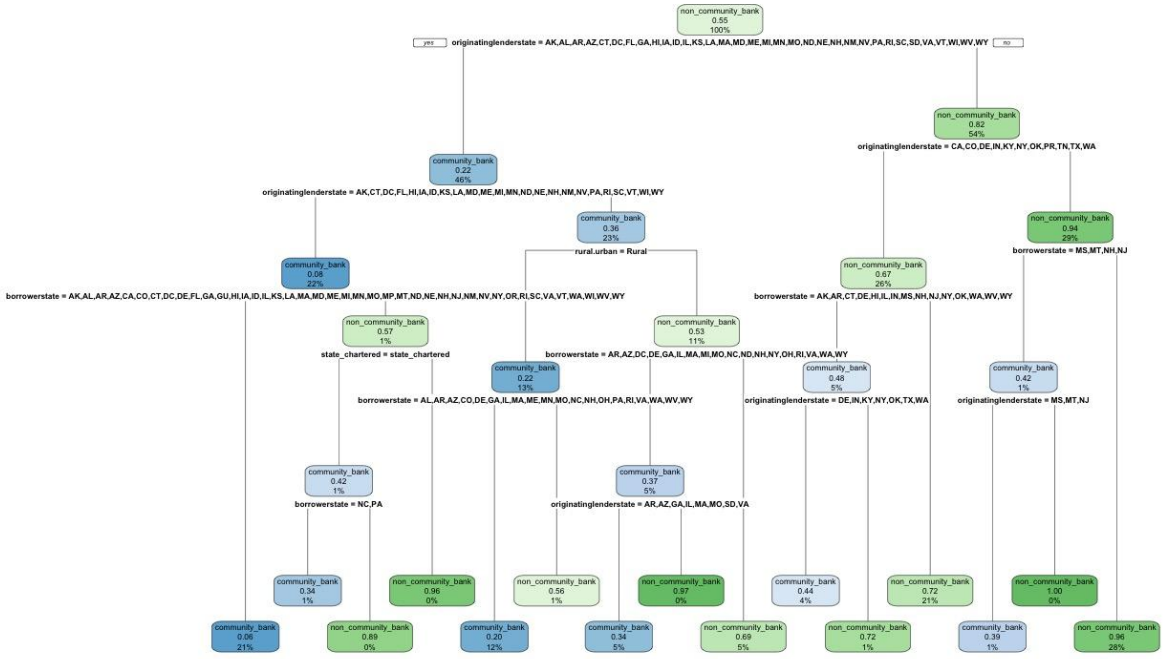


Figure 30

This second decision tree and prediction is established based on the GINI index. It has many leaves and a deeper depth. Table 9 shows different sets of corresponding borrower states and originating lender states below.

Table 9: Prediction II on the distribution of community banks

predicted1	AK	AL	AR	AZ	CA	CO	CT	DC	DE	FL	GA	GU	HI	IA	ID	IL
community_bank	669	3177	3996	192	0	0	228	5	24	779	499	0	206	423	146	958
non_community_bank	0	2261	57	69	11590	443	0	0	34	9	201	12	0	0	0	352

predicted1	IN	KS	KY	LA	MA	MD	ME	MI	MN	MO	MS	MT	NC	ND	NE	NH
community_bank	328	279	30	245	706	112	420	190	595	788	173	154	19	349	374	84
non_community_bank	219	9	164	1	39	2	3	0	6	670	1108	589	754	0	2	0

predicted1	NJ	NM	NV	NY	OH	OK	OR	PA	PR	RI	SC	SD	TN	TX	UT	VA
community_bank	341	126	49	503	6	315	0	1821	0	85	96	76	0	53	13	1065
non_community_bank	8745	1	2	232	3221	513	1427	698	369	0	4	48	411	448	3046	64

predicted1	VT	WA	WI	WV	WY
community_bank	582	1682	6147	596	665
non_community_bank	0	293	3	101	0

For the left nodes, it seems that despite the borrower state which is the same as its originating lender state (about 50%), the states of Alabama, Arkansas, Arizona, California, Colorado, Delaware, Georgia, Guam, Illinois, Massachusetts, Missouri, Northern Mariana Islands, Montana, New Jersey, Oregon, Virginia, Washington, and West Virginia are more likely to be the borrower states instead of lender states. To discover the reason behind it, the prediction of community banks is made. Since Figure 10 provides an insight that the major banks of lender states are community banks, it can be seen that the reason people go to the other states to borrow money is because the states they originally in have very few community banks. However, from the prediction, it shows that the number of community banks in some borrower states such as Virginia, Washington and West Virginia are much higher than the number of non-community banks which might imply that the performance, strategy, or efficiency of community banks in these states are worse than the lender states banks. Policy makers can then establish more policies to help foster the development of community banks in these states and then benefit the ordinary borrowers.

For the right nodes, it focuses more on non-community banks, and there are no overlapping borrower and lender states. Also, the total matches of states is much lower than the community banks, which can be concluded that people prefer to borrow money from urban area's community banks than non-community banks.

c. Decision Tree III:

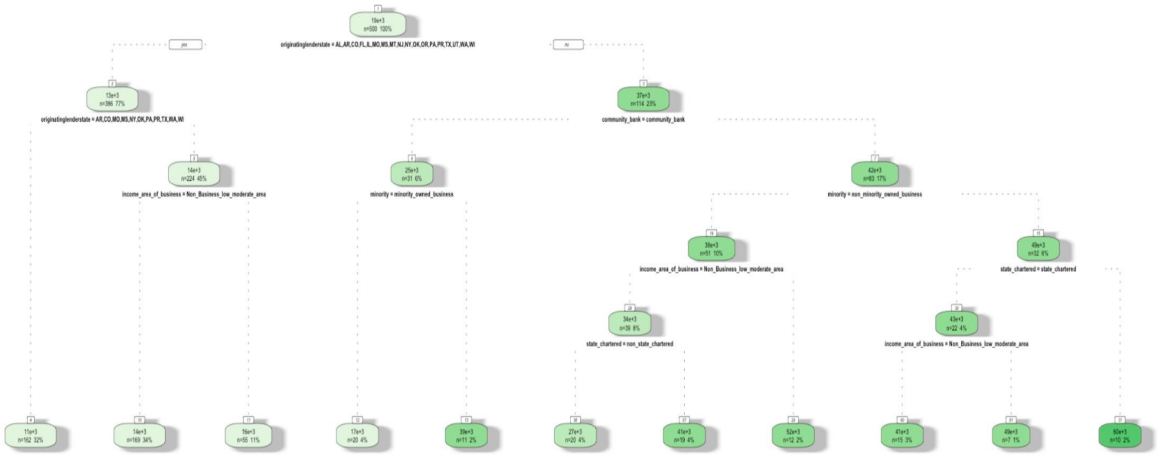


Figure 31

To discover more about the relationship between banks and the approval amount, the third tree *(Figure 31)* is made. This tree shows that for a higher value of approval amount, it mostly happens with the condition of high-income area of originating lender states of Alabama, Arkansas, Colorado, Florida, Illinois, Missouri, Mississippi, Montana, New jersey, New York, Oklahoma, Oregon, Pennsylvania, Texas, Utah, Washington, Wisconsin. For borrowers, they can take advantage and develop a bigger company in these states with a higher borrowing amount. It also shows that for community banks, non-minority owned business have a much higher approval amount than minority owned business which means that government or policy makers can give more financial support or grants to community banks which are only owned by small groups to help them grow.

E. Random Forest:

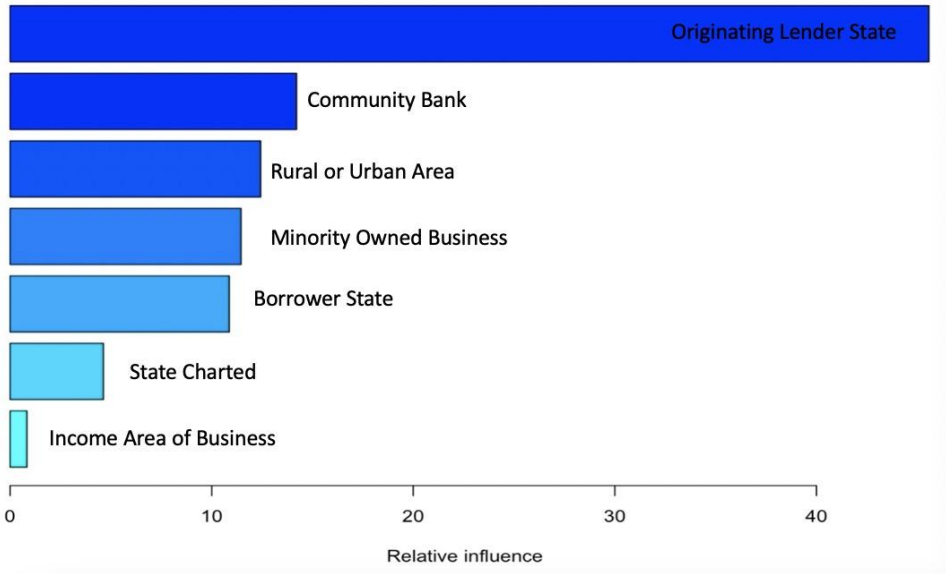


Figure 32

While there is more interpretability in a Decision Tree, Random Forest (ensemble of trees) provides much better prediction accuracy. The Random forest is made to compare the accuracy of prediction of decision trees. Random forest can be applied to a very large amount of dataset. Also, as a combination of hundreds of decision trees, random forest would generate the prediction and result with a higher accuracy value and lower prediction errors, or called RMSE (Root Mean Square Error). This random forest model was found through a grid research of different parameters in each iteration number of trees and the number of parameters to include in each split which will give a different tree each time which decorrelate the ensemble of trees which leads to a better RMSE. The above relative influence of variables shows directly about how variables influenced the prediction of approval amount.

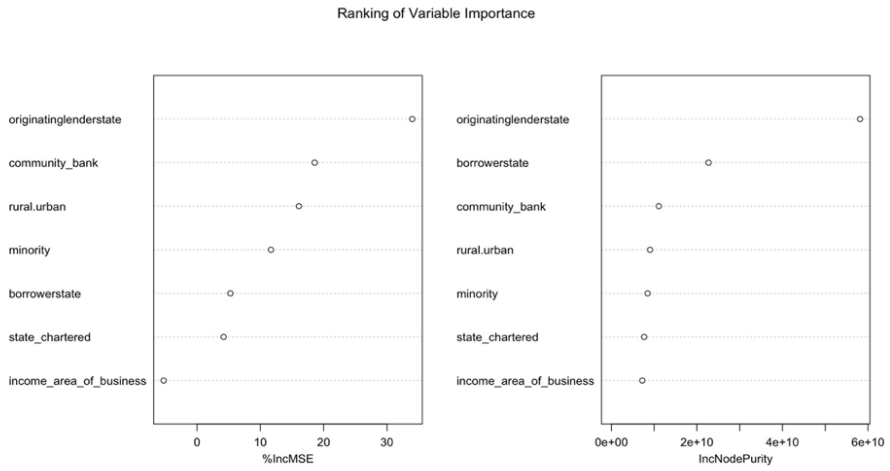


Figure 33

The left plot of Figure 33 is about how much each variable has improved the MSE, which is the most robust and informative measure with the meaning of the higher number, the more important. The right plot is about the NodePurity of the trees which means the more important the variable is, the higher the value is. It shows that the most influential variable is originating lender state, and the second one as community bank, while the least influential variable is income area of business. Policy makers can decide to put more emphasis on these variables if they would like to increase the financial level of banks. It might be explained as different states have their own unique policy of providing approval amounts, and the total amount of value can be very different as if the bank is a community bank. By noticing this, regulators can then research the policy of lender states in the 3rd decision tree model and apply their policy advantages to those states who have a pretty low approval amount. It would help a lot to increase the economic level for all states as a whole.

F. Boosting Models:

Table 10 (Influential variables under Boosting)

Variable	Relative Influence
Originating lender state	45.5814121
Minority	14.2082064
Community_bank	12.4233543
State_chartered	11.4555700
Income_area_of_business	10.8694100
Borrowerstate	4.6265035
Rural.urban	0.8355438

To decide the most accurate model, the boosting model is tested as a comparison of decision tree and random forest. Boosting uses sets of weak learners (small trees) to correct the residuals slowly using a learning rate. The process of generating a boosting model is firstly selecting a random sample of data and then fitted and trained sequentially to a more modified version of prediction. Mentioned By by Gareth James(*"An Introduction to Statistical Learning"*, page 352) [3], it is known that since each new tree is fit to the signal that is left over from the earlier trees, and shrunken down before it is used, trees are growing very slowly and successively in boosting model.

For the result table generated under the Boosting model, the variable influence is different than before as the minority owned business is more influential than the community bank. To fairly compare the utility of decision tree, random forest, and boosting model and

choose the best model for giving suggestions to policy makers, regulators, and ordinary borrowers, a summary table of RMSE value is created as follows.

Table 11: Summary Table of RMSE

Machine Learning Methods	RMSE Value
Decision Tree	384945.0923
Random Forest	20707.474665
Boosting Model	59148.19

The Summary table shows Root Mean Square Error of all three models. Since the test RMSE provides a better understanding of the model performance when unseen data (by the model) is fed into the model, it is a vital measure for prediction analysis of machine learning models. Clearly, Random Forest owns the lowest error, and would be the best model to choose. Therefore, it validates the importance of community bank and lender state among all variables to the amount approved by a bank. Policy makers can then focus more on giving more financial support to states with lower approval amounts and expand the business range of community banks.

6. Results & Future Steps (Neural Networks)

By doing hypothesis testing, it could be roughly suggested that no matter in rural or urban areas, and any level of income regions, both community and non-community banks might have a large forgiveness amount. For most originating lender states, there is no obvious relationship between forgiveness amount level and whether the business is located in a rural or in an urban area. Similar results hold for lenders within a low to moderate- or a high-income area, but there are still some states that don't follow the same pattern. The PPP loan forgiveness amount of community banks in rural areas is higher than non-community banks, however, this pattern doesn't apply to urban areas.For regression analysis, it could be concluded from the final models that banks with a high forgiveness ratio and employees ratio are more likely to have a high approval ratio. And community banks and urban banks are more likely to get a higher approval ratio, but state chartered banks are less likely to get a higher approval ratio. In addition, urban banks and banks with a higher employees ratio are less likely to obtain a higher forgiveness ratio. After comparing the RMSE value of Decision Tree, Random Forest, and Boosting Modeling, the prediction of Random Forest was proved to be the most accurate. It can be concluded that under the Random Forest, the community bank indeed plays a vital role in the financial power of a state. Under the decision tree, it is also shown that for urban areas of specific states, the approval amount is very high which can give regulators the reason that there are many people going to other states to borrow money.

Combined with the results of Machine Learning models of Random Forest, the relationship between forgiveness, approval, employment amount and the community banks as well as lender states are discovered. The Deep Learning method (Neural Networks) can be

applied to increase the model prediction on performance of community banks. Neural Networks is an algorithm of analyzing multiple layers of variables and predictions are made depending on the weight of each layer. Moreover, since the influential variables are limited for the dataset provided, if more datasets or variables such as fraud count and credit card transactions in years are researched, Neural Networks can also be applied to find fraud detection and score the risks of lending money to a borrower for each community bank's regulators; and therefore offer state policy makers a better understanding of the performance of community banks.

7. Conclusion & Recommendations

a. Although some states, such as Virginia, Washington and West Virginia have a relatively large number of community banks, people in these states still prefer to go to other states to borrow money. Policymakers can therefore be suggested by analyzing the advantage of lender states' policy of community banks and setting up more financial support to solve the long distance borrowing problem.

b. Providing more programs on the federal, state, county and city levels which can support and foster minority business enterprises to grow, and can also mitigate the monopoly from large business.

c. Making more regulations and offering benefits to community banks to fill the gap between low approval amount and high approval amount and then balancing the distribution(Same for forgiveness amount).

d. Some insights can be given for policy makers that since the average ratio of forgiveness, employment, and approval rates might have some potential relationships with each other in one state, if a government would like to increase the capability of employment ratio, it can be suggested to increase the financial power first. However, this point needs to be argued before enacting any policies.

Appendix

All Codes/ Datasets/ Extra Plots are committed on Github:

https://github.com/zw304/CSBS_Data_Analytics_Competition.git

Reference

[1] WebFOCUS 8 Technical Library

https://infocenter.informationbuilders.com/wf80/index.jsp?topic=%2Fpubdocs%2FRStat16%2Fsource%2Ftopic49.htm

[2] Business Loan Program Temporary Changes; Paycheck Protection Program

https://www.federalregister.gov/documents/2020/04/15/2020-07672/business-loan-program-temporary-changes-paycheck-protection-program

[3] *An Introduction to Statistical Learning(2nd Edition),* Gareth James, page 352 , August 4th, 2021

[4] Menard S. Applied Logistic Regression Analysis(2nd edition), SAGE Publications, Inc; 2001.

Appendix

Codes: commited on Github https://github.com/zw304/CSBS_Data_Analytics_Competition.git

Table 12 on finding Best “mtry” of Random Forest

Mtry	Forest.MSE
1	558723177
2	446950193
3	465282929
4	469766478
5	469401987
6	438223793
7	458981150
8	460754572
9	454014925
10	466707256

Table 13 on Accuracy Checking

	Community bank	Non-community bank	Mean Decrease Accuracy	Mean Decrease Gini
State_chartered	3.598888	-1.299449	1.279108	2.401664
Borrower state	22.242294	-9.563389	12.538251	17.395547
Originating lender state	50.453486	16.377035	48.368445	46.928187
Rural.urban	10.152075	-5.346130	4.462529	4.784463
Business_low_to_moderate	-5.133432	-6.185049	-7.853851	2.652367

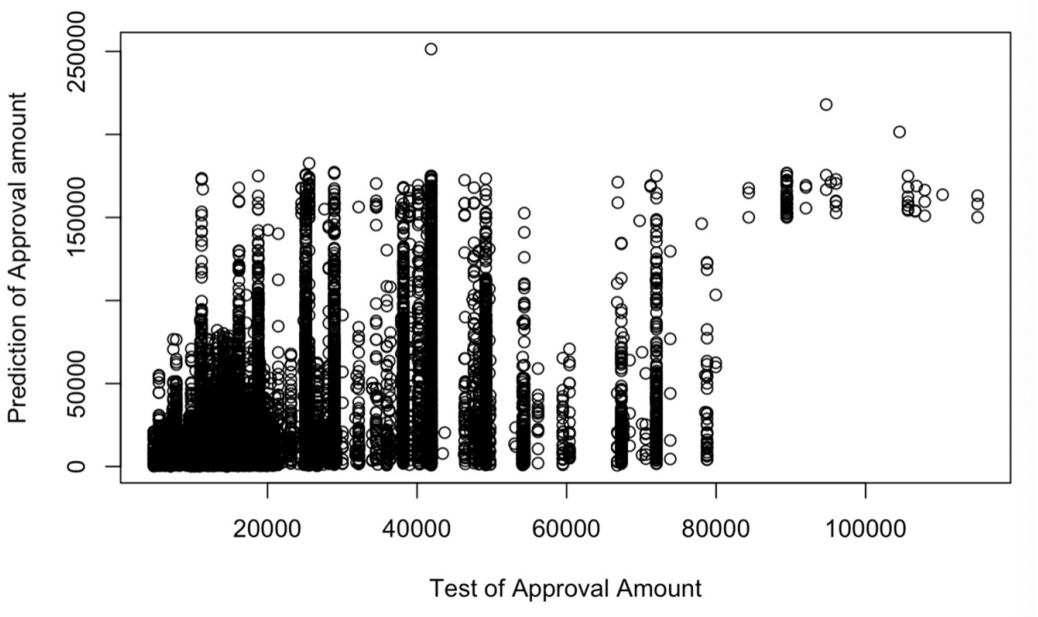


Figure 34 on Approval Amount Prediction