SMU | Lyle School of Engineering

The Role of Community Banks

During the Pandemic

Final Report

Competition Members

Melissa Serrano

Annalise Sumpon

Madi Tedrow

Professor

Dr. Eli Olinick

April 24, 2022

Table of Contents

List of Figures and Tables

1.0 Executive Summary

The Data Analytics competition hosted by Conference of State Bank Supervisors' (CSBS) is focused on conducting economic research to better understand community banking through data analytics.

For the competition, our team decided to examine Paycheck Protection Program (PPP) data to determine the role of community banks in the distribution of the loans. The Paycheck Protection Program was developed to help struggling small businesses during the COVID-19 pandemic. Community banks specifically played a critical role in disseminating loans to vulnerable businesses and entities since they primarily serve small counties across the US. Our team hypothesized that there were statistically significant variables such as business type, size, and ownership that could be used in determining whether a business received a PPP loan from a community bank vs a non-community bank.

By implementing Logistic Regression, Classification and Regression Tree (CART), and Naive Bayes data models, we discovered that variables related to business ownership, business type, business location, and state chartership were key to determining whether a loan originated from a community or a non-community bank. Correspondingly, we were able to determine that business size did not play a significant role in distinguishing community banks from non-community banks in the distribution of PPP loans. Our findings ultimately suggest a relationship between the aforementioned key categorical variables and community banks that could be implemented as a framework in the future by stakeholders to promote a more equitable and robust financial system in and out of times of crisis.

2.0 Problem and Background

With the onset effect of the COVID-19 pandemic in the midst of 2020, many small businesses fell victim to economic hardship. In an effort to aid small businesses, the Small Business Administration (SBA) set up the Paycheck Protection Program (PPP). The PPP was designed to help small businesses keep workers employed during the pandemic by providing funds through short-term loans backed by the SBA. During the COVID-19 pandemic, community banks played a major role in distributing Paycheck Protection Program (PPP) loans. Preliminary research conducted by the Conference of State Bank Supervisors (CSBS) shows that community banks were at the forefront in providing these needed funds to small businesses. Community banks had a disproportionately greater share in lending to small businesses than their larger and more complex counterparts. However, the question lies in the nuances of the impact that community banks play with the distribution of PPP loans during the COVID-19 pandemic. In other words, is it possible that there are specific communities or business characteristics that were affiliated with the role of community banks with PPP loans?

In order to further our understanding to help scholars, policymakers, and the public better understand the role of community banks during the COVID-19 pandemic, we utilized the Paycheck Protection Program data provided by CSBS which is a smaller and cleaner version of the data provided by the SBA. Since our goal is reliant on gaining further insights regarding various business characteristics, such as race or ethnicity, we merged a subset of data from the original SBA dataset to the CSBS dataset to expand our data's scope. Our goal is to determine if there exists statistically significant differences or patterns in the business type, business size (number of jobs), and business ownership (e.g., racial minority or gender) for the businesses that received the PPP loans within community banks and/or non-community banks.

2.1.1 PPP Loans

The Paycheck Protection Program (PPP) was established by the CARES Act and implemented by the Small Business Administration (SBA) during the COVID-19 pandemic. Billions of dollars worth of low interest loans were given to small businesses in order to provide resources needed to maintain payroll, rehire laid-off employees, support resource expenditure etc. Interestingly enough, a large portion of these loans are being forgiven, meaning that businesses do not have to pay them back; however, this raises concern because many people worry that the loans were not used in the way they were intended to be used.

While the PPP loans were meant to help "businesses keep their workforce employed during the COVID-19 crisis," according to the U.S. Small Business Administration, it is estimated that only 23 to 34 percent of PPP dollars were used towards directly helping the workers. PPP loans could also be used to pay interest on mortgages, rent, and utilities which may be where the money may have gone instead. Because community banks tend to have a closer connection with their customers, community banks may have played a larger role in distributing PPP loans focused on workers rather than non-community banks.

Additionally, there was concern over who PPP loans were distributed to and headlines across a variety of news sources began to reflect this concern over the years since 2020:

- "Did Asia- and American Indian-Owned Businesses Receive Their Fair Share of PPP Loans?" *(Heartland Forward- August 11, 2020)*
- "The Paycheck Protection Program Failed Many Black-owned Businesses" *(Vox- October 5, 2020)*
- "Minority Entrepreneurs Struggled to Get Small-Business Relief Loans" *(New York Times- April 4, 2021)*
- "Racial Bias Affected Black-owned Small Businesses Seeking Pandemic Relief Loans, Study Finds" *(The Washington Post- October 15, 2021)*

While it is not possible to go back and fix the issues that were discovered from the distribution of the PPP loans, it is possible to learn from the experience and apply it to future situations. Although the distribution of the loans may not have been equitable, our analysis may allow for community banks to be identified and classified more readily in the future to get loans to communities in need.

2.1.2 Community Banks

2.1.2.1 Critical Function within Counties

Besides distributing PPP loans, community banks serve a critical role in smaller and rural markets across the US by providing credit, liquidity, and investments to communities. These types of banks are traditionally headquartered in counties with populations of 50,000 or less.

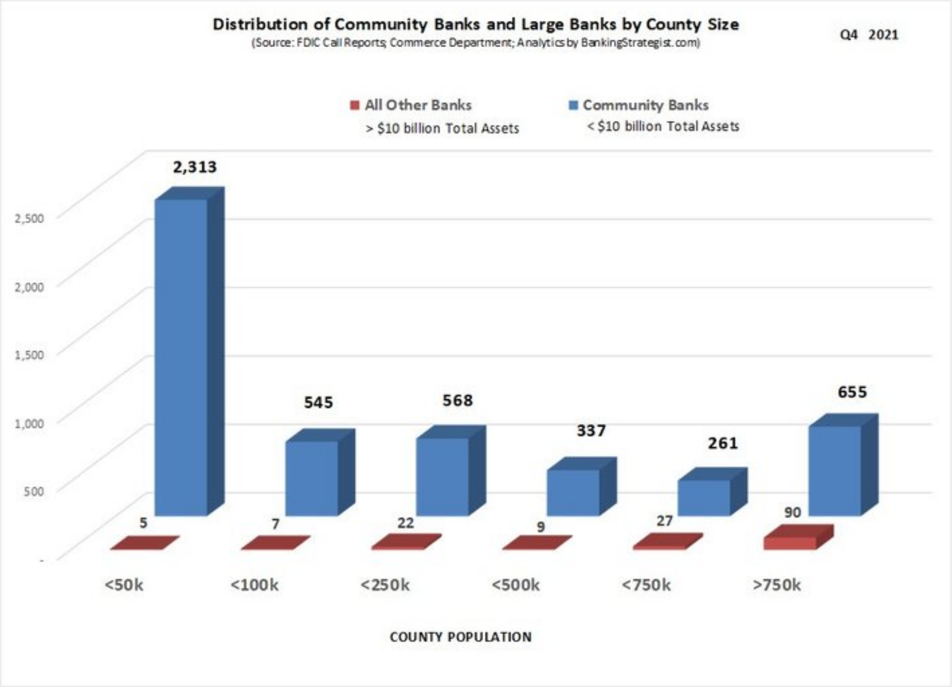
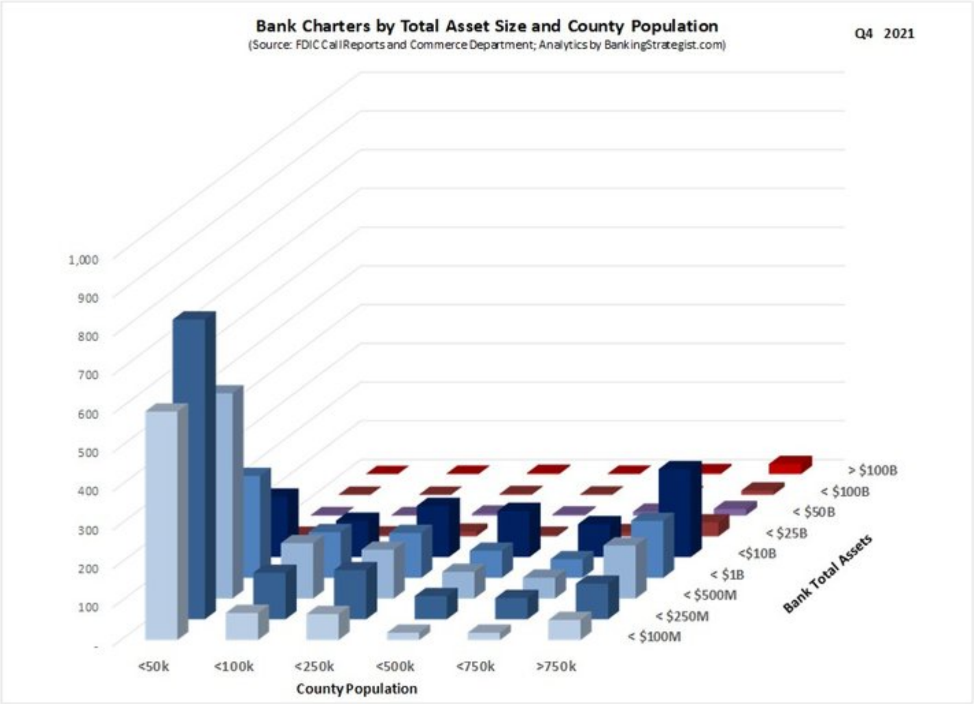


Figure 1: Distribution of community banks and large banks by county size



Bank Charters by Total Asset Size and County Population

(Source: FDIC Call Reports and Commerce Department; Analytics by BankingStrategist.com)

Q4 2021

Figure 2: Bank charters by total asset size and county population

Reflective of the smaller population size, smaller community banks are likely the ones that serve these kinds of markets and regardless of size, community banks remain the predominant financial institution in counties with smaller populations. Figures 1 and 2 illustrate, 2313 community banks are headquartered in counties with populations less than 50,000 and serve approximately 28 million people across 1400 US counties with populations of less than 50,000.

2.1.2.2 FDIC Designation

The FDIC designates a bank as a community bank by using a five-step protocol. First, all charter-level data under each holding company is aggregated into a single organization. Next, certain banks are filtered out through exclusion if the bank holds more than 50% of total assets in specialty banks such as credit card specialists, consumer nonbanks, industrial loan companies, trust companies, and bankers' banks. Certain banks are then included based on if they engage in basic banking activities as measured by the total loans-to-assets ratio (greater than 33 percent) and the ratio of core deposits to assets (greater than 50 percent). Organizations that operate within a limited geographic scope are also included in the aggregation based on a minimum and maximum number of total banking offices. Lastly, the FDIC determines an asset-size limit which is also adjusted upward over time where the limits on banking activities and geographic scope are waived.

2.1.3 Legislation Relevancy

Determining what variables correlate with community bank loans is important because it could allow key stakeholders (such as CSBS, SBA, regulators, policymakers, and banks) to proactively identify specific banks in need of financial aid. Our analysis provides a clear and quantitative way to substantiate and support claims on the impact of financial downwinds on minority and low-income populations. Even further, stakeholders could take steps to enable and strengthen the communities that community banks serve- those with smaller populations in more rural areas with lower income and higher minority populations. Specifically, stakeholders could take steps to further investigate our key variables for community bank identification, create special designation labels for banks with community bank characteristics, and expand credit willingness for community banks to receive loans as prevention against financial downwind.

3.0 Data

The Paycheck Protection Program data is provided by CSBS which is a smaller and cleaner version of the data provided by the Small Business Administration (SBA) focused on community banks.

3.1 Data Preparation and Cleaning

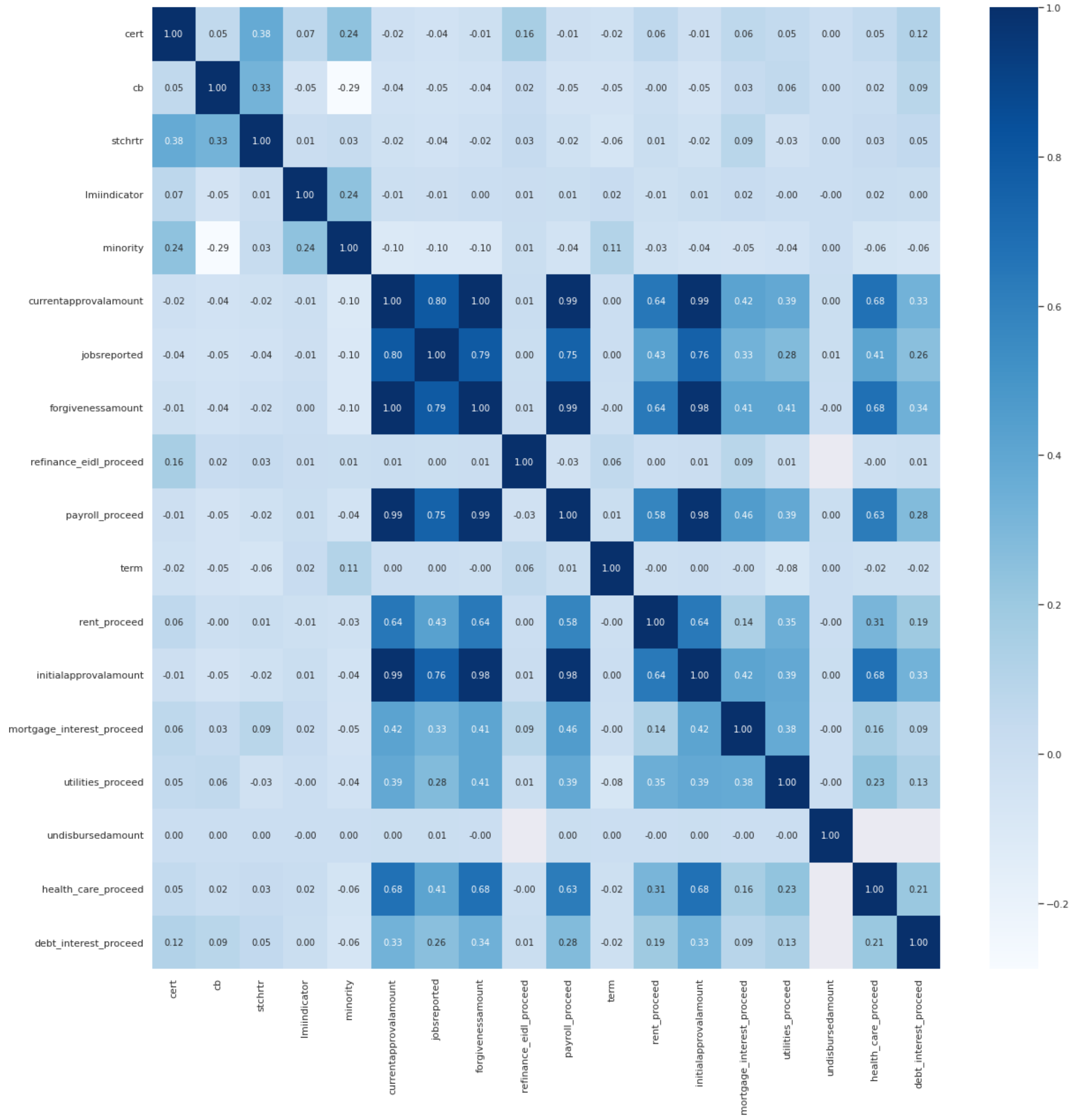
The CSBS dataset is a subset of the PPP data extracted from the SBA's full dataset. The provided set was of size 8,614,374 observations with 19 variables. In order to merge the CSBS dataset with the original SBA PPP plus 150K dataset, we merged by using a left-join on the ‘loannumber’ identification number to combine additional variables for further insight into our goal from the original SBA dataset. To avoid possible duplication we joined CSBS data and SBA data using *loannumber* and *borrowerstate* to ensure that we were expanding CSBS data with correct values. After merging new SBA columns to the CSBS data, there were a total of 58 variables with the same size of 8,614,374 observations.

In order to prepare our dataset for modeling, we needed to extract needed columns and fill or drop null values. To do this, in order to narrow down our columns, we extracted columns that pertained to our goal of business characteristics. Additionally, we compared the amount of missing data per column and removed columns that had more than two-thirds of the data missing/null.For quantitative variables, all rows containing missing values were removed from the dataset. For categorical variables, such as ‘race’ or ‘businesstype’, the remaining null values were adjusted using a common practice of replacing the null data point with the mode of each column. Our resulting dataset, after data cleaning and removing invalid values, contains 15 variables with 122,155 observations. Holistically, this resulted in a prepared reliable dataset that combined both the provided CSBS data and additional SBA data for modeling.

3.2 Variable Selection

In order to proceed with modeling the full cleaned dataset, it is important to get a general understanding of the different variable relationships. Below, Figure 3 displays a correlation matrix from the uncleaned dataset to aid our understanding into variable selection for our cleaned dataset.

Based on the correlation matrix, we can see that *initialapprovalamount*, *forgivenessamount*, and *jobsreported* with *currentapprovalamount* have high correlation. This is important to note for future modeling. It is also worth noting that *currentapprovalamount* is also highly correlated with *payroll_proceed*. These specified variables may show to be crucial for predicting loan distribution between community banks and non-community banks. Highly correlated variables such as those identified will be removed as needed when refining our models to address any redundancy in our results.



	cert	cb	stchrtr	lmiindicator	minority	currentapprovalamount	jobsreported	forgivenessamount	refinance_eidl_proceed	payroll_proceed	term	rent_proceed	initialapprovalamount	mortgage_interest_proceed	utilities_proceed	undisbursedamount	health_care_proceed	debt_interest_proceed
cert	1.00	0.05	0.38	0.07	0.24	-0.02	-0.04	-0.01	0.16	-0.01	-0.02	0.06	-0.01	0.06	0.05	0.00	0.05	0.12
cb	0.05	1.00	0.33	-0.05	-0.29	-0.04	-0.05	-0.04	0.02	-0.05	-0.05	-0.00	-0.05	0.03	0.06	0.00	0.02	0.09
stchrtr	0.38	0.33	1.00	0.01	0.03	-0.02	-0.04	-0.02	0.03	-0.02	-0.06	0.01	-0.02	0.09	-0.03	0.00	0.03	0.05
lmiindicator	0.07	-0.05	0.01	1.00	0.24	-0.01	-0.01	0.00	0.01	0.01	0.02	-0.01	0.01	0.02	-0.00	-0.00	0.02	0.00
minority	0.24	-0.29	0.03	0.24	1.00	-0.10	-0.10	-0.10	0.01	-0.04	0.11	-0.03	-0.04	-0.05	-0.04	0.00	-0.06	-0.06
currentapprovalamount	-0.02	-0.04	-0.02	-0.01	-0.10	1.00	0.80	1.00	0.01	0.99	0.00	0.64	0.99	0.42	0.39	0.00	0.68	0.33
jobsreported	-0.04	-0.05	-0.04	-0.01	-0.10	0.80	1.00	0.79	0.00	0.75	0.00	0.43	0.76	0.33	0.28	0.01	0.41	0.26
forgivenessamount	-0.01	-0.04	-0.02	0.00	-0.10	1.00	0.79	1.00	0.01	0.99	-0.00	0.64	0.98	0.41	0.41	-0.00	0.68	0.34
refinance_eidl_proceed	0.16	0.02	0.03	0.01	0.01	0.01	0.00	0.01	1.00	-0.03	0.06	0.00	0.01	0.09	0.01		-0.00	0.01
payroll_proceed	-0.01	-0.05	-0.02	0.01	-0.04	0.99	0.75	0.99	-0.03	1.00	0.01	0.58	0.98	0.46	0.39	0.00	0.63	0.28
term	-0.02	-0.05	-0.06	0.02	0.11	0.00	0.00	-0.00	0.06	0.01	1.00	-0.00	0.00	-0.00	-0.08	0.00	-0.02	-0.02
rent_proceed	0.06	-0.00	0.01	-0.01	-0.03	0.64	0.43	0.64	0.00	0.58	-0.00	1.00	0.64	0.14	0.35	-0.00	0.31	0.19
initialapprovalamount	-0.01	-0.05	-0.02	0.01	-0.04	0.99	0.76	0.98	0.01	0.98	0.00	0.64	1.00	0.42	0.39	0.00	0.68	0.33
mortgage_interest_proceed	0.06	0.03	0.09	0.02	-0.05	0.42	0.33	0.41	0.09	0.46	-0.00	0.14	0.42	1.00	0.38	-0.00	0.16	0.09
utilities_proceed	0.05	0.06	-0.03	-0.00	-0.04	0.39	0.28	0.41	0.01	0.39	-0.08	0.35	0.39	0.38	1.00	-0.00	0.23	0.13
undisbursedamount	0.00	0.00	0.00	-0.00	0.00	0.00	0.01	-0.00		0.00	0.00	-0.00	0.00	-0.00	-0.00	1.00		
health_care_proceed	0.05	0.02	0.03	0.02	-0.06	0.68	0.41	0.68	-0.00	0.63	-0.02	0.31	0.68	0.16	0.23		1.00	0.21
debt_interest_proceed	0.12	0.09	0.05	0.00	-0.06	0.33	0.26	0.34	0.01	0.28	-0.02	0.19	0.33	0.09	0.13		0.21	1.00

Figure 3: Correlation Matrix on Merged Dataset

3.2.1 Chi-Squared Test of Independence for Feature Selection

In order to explore the data and choose features methodically, we conducted a Chi-Squared amongst categorical data types to narrow down relevant features for building our models. The Chi-Square test of independence is a statistical test to determine if there is a significant relationship between 2 categorical variables. In other words, the Chi-Square statistic is used to test if there is a significant difference in the observed vs the expected frequencies of both variables. In our case, if the p-value is less than the critical value- level of significance 0.05, we reject the null hypothesis and believe that the variables are associated with community banks.

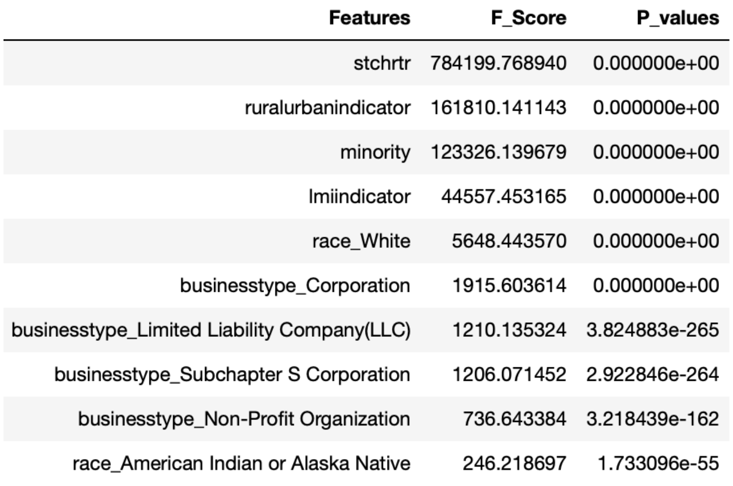
$$\chi = \sum \frac{(Observed - Expected)^2}{Expected}, \; Expected = \frac{RowTotal \times ColumnTotal}{OverallTotal}$$

H_o: $No\ association\ with\ Community\ Banks$

H_a: $There\ is\ evidence\ of\ association\ with\ Community\ Banks$

We implemented the Chi-squared test on various categorical data types within the data set to pull out most significant features. Table 1 shows the features with highest significance using Chi-Squared Test of Independence.

Table 1: Chi-Squared Method for Categorical Feature Selection



Features	F_Score	P_values
stchrtr	784199.768940	0.000000e+00
ruralurbanindicator	161810.141143	0.000000e+00
minority	123326.139679	0.000000e+00
lmiindicator	44557.453165	0.000000e+00
race_White	5648.443570	0.000000e+00
businesstype_Corporation	1915.603614	0.000000e+00
businesstype_Limited Liability Company(LLC)	1210.135324	3.824883e-265
businesstype_Subchapter S Corporation	1206.071452	2.922846e-264
businesstype_Non-Profit Organization	736.643384	3.218439e-162
race_American Indian or Alaska Native	246.218697	1.733096e-55

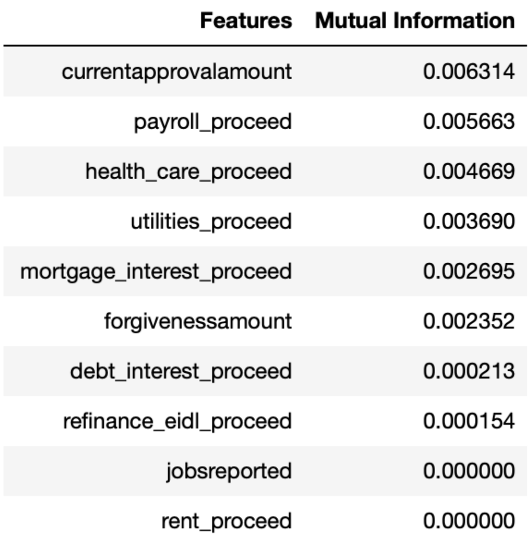
3.2.2 Mutual Information Feature Selection

To study the significance of quantitative variables within the data set, we decided to perform a Mutual information test between quantitative features. Mutual information (MI) measures the reduction in uncertainty for one variable given a known value of the other variable to estimate the contribution of a variable towards another variable. This method detects how much a target variable is impacted if a feature is added or removed, which is a positive way to create relevant models. This statistic represents the entropy of each variable, which measures or quantifies the amount of information obtained about one random variable, through the other random variable. Additionally, by identifying MI, we are able to select variables that maximize the information gain, which minimizes the joint entropy. By using this method for the purpose of feature selection, we are able to choose features that maximize MI to ensure relevancy. The statistical valuation builds off of the concept of joint entropy as can be seen below:

- $H(A,B) \;=\; -\sum_{i,j} p(i,j) \times log[\,p(i,j)] \quad (joint\ entropy)$
- $MI(A,B) \;=\; \sum_{x}\sum_{b} p(x,b) \;\times\; log(\frac{p(x,y)}{p(x)p(y}) \;\leq\; 1$

For our study, it is a prime method to use for the purposes of reducing the amount of quantitative variables in our dataset, because it prepares our dataset to detect non-linear relationships and is effective for both regression and classification modeling techniques. Mutual Information scores a feature between 0 and 1, thereby indicating if both the variables are deterministic of community banks. Table 2 shows the feature results of the Mutual Information method.

Table 2: Mutual Information Method for Quantitative Feature Selection



Features	Mutual Information
currentapprovalamount	0.006314
payroll_proceed	0.005663
health_care_proceed	0.004669
utilities_proceed	0.003690
mortgage_interest_proceed	0.002695
forgivenessamount	0.002352
debt_interest_proceed	0.000213
refinance_eidl_proceed	0.000154
jobsreported	0.000000
rent_proceed	0.000000

3.3.3 Final Variable Selection

Our goal is to determine the statistical differences and patterns between the business characteristics, specifically variables that indicate community metrics and loan metrics. Keeping our goal in mind, as well as results from both Chi-squared and Mutual Information test, we compiled significant variables that will be used for modeling. Below, Table 3 identifies the following variables as points of interests for our data analysis:

Table 3: Variable Field Names, Definitions, and Values

Field Name	Definition and Values
cb	0 = Non-Community Bank 1 = Community Bank
stchrtr	0 = Non-State Chartered 1 = State Chartered

ruralurbanindicator	U = Urban
	R = Rural

lmiindicator	0 = Business is not in low-to-moderate income area 1 = Business is in Low-to-moderate income area
currentapprovalamount	Loan amount approved
jobsreported	Number of jobs the business has reported
forgivenessamount	Loan amount forgiven
race	Borrower Race Description ['White', 'Asian', 'American Indian or Alaska Native', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'Multi Group', 'Puerto Rican', 'Eskimo & Aleut']
minority	0 = No minority owned business 1 = Yes minority owned business
ethnicity	Borrower Ethnicity Description 'Unknown/NotStated' 'Not Hispanic or Latino' 'Hispanic or Latino'
businesstype	Business Type Description ['Sole Proprietorship', 'Limited Liability Company(LLC)', 'Corporation', 'Subchapter S Corporation', 'Non-Profit Organization', 'Joint Venture', 'Limited Liability Partnership', 'Partnership', 'Trust', '501(c)6 – Non Profit Membership', 'Cooperative', 'Independent Contractors', 'Professional Association', 'Tribal Concerns', 'Self-Employed Individuals', '501(c)3 – Non Profit', 'Employee Stock Ownership Plan(ESOP)', 'Non-Profit Childcare Center', 'Single Member LLC', 'Tenant in Common', 'Housing Co-op', '501(c) – Non Profit except 3,4,6,', '501(c)19 – Non Profit Veterans', 'Rollover as Business Start-Ups (ROB']
payroll_proceed	Amount of proceeds that were reported to be applied towards Payroll costs

3.3 Model Selections

As a team, we looked at several different models to analyze the merged SBA data and considered the pros and cons of each in preliminary modeling.

- *Linear regression* gives information on the significance of each variable, but generates low accuracy in preliminary modeling. This may have been due to our dataset having a combination of variable types with the majority being categorical variables taking a 0 or 1 binary value. Because our target variable is classification based as well (community bank or non-community bank), linear regression would likely have had higher accuracy if our dataset had more continuous variables.
- *Logistic regression* is used when the dependent variable is binary in nature and supports categorization into discrete classes. This type of model should have good accuracy when the dataset is linearly separable and can interpret model coefficients as indicators of feature importance. However, it can be difficult to obtain complex relationships and the target variable of the logistic regression is bound to being a discrete variable (which was not a problem in our case).
- *Polynomial regression with multiple variables* should work well on any size dataset and on non-linear problems but the correct polynomial degree needs to be chosen for tradeoffs between bias/variance. In preliminary testing, we found this model had higher accuracy than linear regression but comparatively low accuracy compared to other models. This may have been due to the presence of outliers affecting the performance or the lack of a curvilinear relationship between the dependent variable (community bank) and selected independent variables.
- *Classification and regression tree (CART) models* are powerful with classification and prediction for class labels. A decision tree is constructed with understandable rules and can handle continuous and categorical variables. The decision tree also provides a clear understanding of what variables are most important for prediction or classification based on the selected splitting criteria; however, they can be difficult and timely to prune for optimal weights and splitting fields.
- *Extreme gradient boosting (XGBoost)* relies on multiple decision trees to determine the final output with weights gradually assigned to trees based on their accuracy scores. As a result, this technique can be used to run cross-validation after each iteration but this comes at a time cost because every classifier is forced to address the predecessor's errors.

- *K-nearest neighbors (KNN)* is applicable in many real-life classification situations such as with PPP data because it is non-parametric and does not make underlying assumptions about the data. Despite this, in our preliminary modeling, we found it needed a high number of neighbors and the results were not easily interpretable.
- *K-means clustering* is a useful tool for categorizing items into groups with the unsupervised k-means algorithm. The algorithm can be used on unlabeled datasets (where there are unknown clusters) and with non-linearly separable data. On the downside, we found in our variable analysis that some of our numerical variables were too correlated so the algorithm struggled to identify a number of *k* clusters to advance.
- *Naive Bayes* is a machine learning algorithm that uses Bayes' Theorem and assumes all predictors are independent, then assigns a level of importance to each feature to generate a classification model based on frequency and probability. This model works well for multi-class prediction problems and requires less training data if its assumptions of independence hold true. The algorithm does face the risk of a "zero-frequency problem" where it assigns zero to a categorical variable but this can be overcome with a smoothing technique.

For our goal specifically, we aimed to conduct different modeling approaches that exhibited supervised machine learning techniques. Supervised learning can be separated into two types of problems: classification and regression. As a result, after weighing the pros and cons of our preliminary modeling along with considerations for scalability, we chose to perform three modeling techniques: Logistic Regression, Classification And Regression Tree (CART), and Naive Bayes. Besides the advantages of the models listed above, using a combination of models allows us to validate our results in multiple ways and prevent bias in our final results. Our goal is to predict community banks versus non-community banks using our features across three different modeling techniques to gain an understanding and validate community banks' impact when distributing PPP loans.

4.0 Modeling

Based on our model section, we chose to conduct three models: Logistic Regression, Classification and Regression Trees, and Naive Bayes to answer our goal of what role community banks played in PPP loan distribution.

4.1 Logistic Regression

Logistic regression models are used to purposefully understand relationships through multiple features to solve binary classification problems. In this approach, logistic regression is utilized to predict binary values 0 or 1 for whether a PPP loan was distributed from a non-community bank or community bank.

The regression modeling algorithm is based on the following equation:

$$P = \frac{1}{1+e^{-(\beta_0+\beta_0 x)}}$$

For this logistic regression, we used community metrics and loan metrics as inputs to create a final equation. The final equation is created using weighted coefficients that give insight into which features are most important to predicting if loans were distributed by community banks. By identifying high-weighted coefficients, these would indicate which specific community metrics are heavily impacted by community banks. Below Figure 4 depicts the bimodal density curve of community banks within the datasheet, where it depicts an expected normal distribution of each peak. When using a logistic regression, the bimodal distribution gives insight on the seemingly balanced nature of the data towards non-community banks and community banks, and how we should expect our classification model to perform.

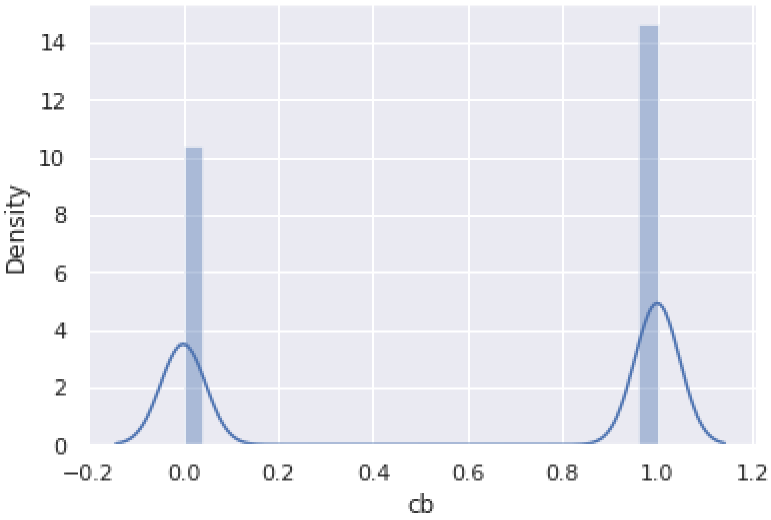
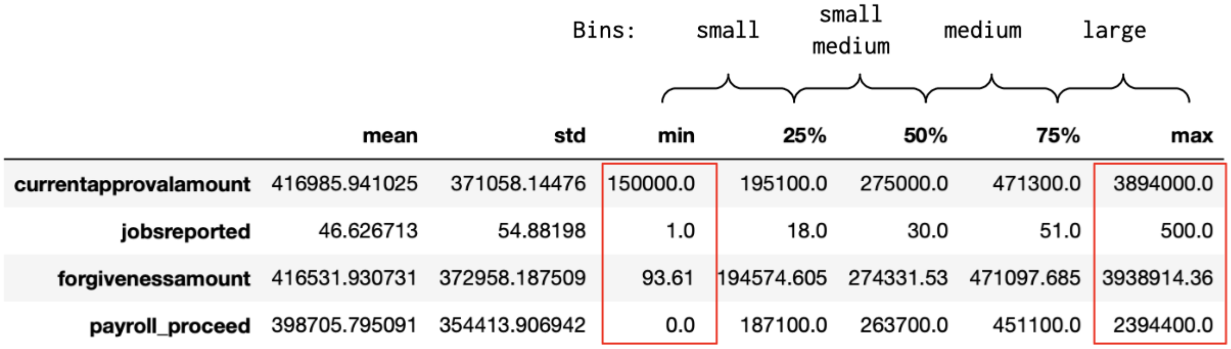


Figure 4: Density curve visualization to predict cb or non-cb

The key characteristic of Logistic regression is that it is able to use categorical inputs to describe create and purposeful logistic equations for classification. Due to this, it is necessary to categorize our quantitative features, such as *currentapprovalamoun*t, *forgivenessamount*, *jobsreported*, and *payroll_proceed*. The method of binning was used by examining the statistical summary of the quartiles as categories.

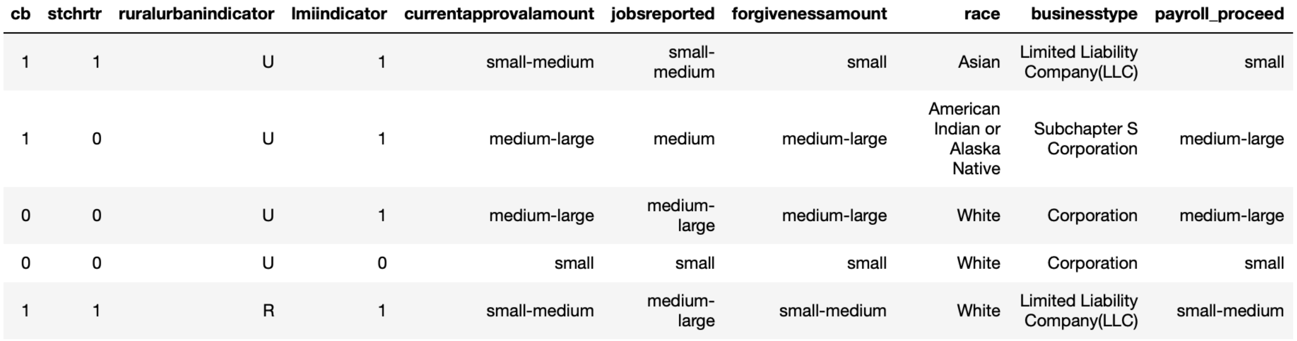
Table 4 explores the binning method and depicts a sample of five of our binning methods that is used for our logistic regression model.

Table 4: Binning Numeric Data for Logistic Regression



Bins: small | small medium | medium | large

	mean	std	min	25%	50%	75%	max
currentapprovalamount	416985.941025	371058.14476	150000.0	195100.0	275000.0	471300.0	3894000.0
jobsreported	46.626713	54.88198	1.0	18.0	30.0	51.0	500.0
forgivenessamount	416531.930731	372958.187509	93.61	194574.605	274331.53	471097.685	3938914.36
payroll_proceed	398705.795091	354413.906942	0.0	187100.0	263700.0	451100.0	2394400.0

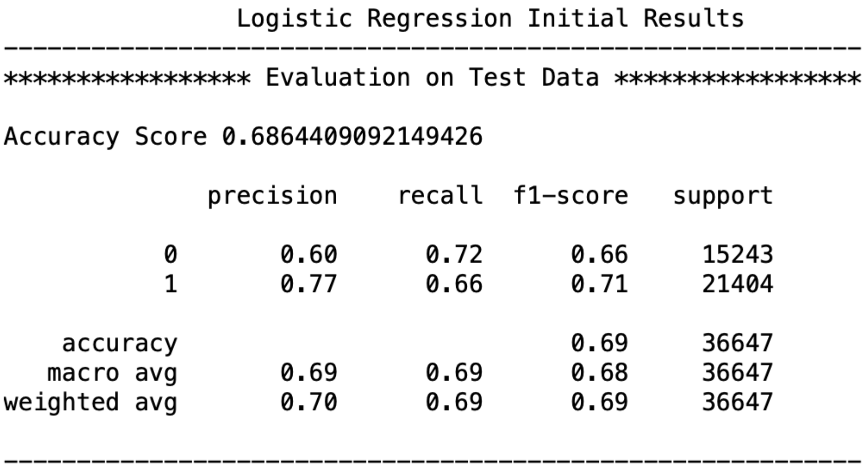


cb	stchrtr	ruralurbanindicator	lmiindicator	currentapprovalamount	jobsreported	forgivenessamount	race	businesstype	payroll_proceed
1	1	U	1	small-medium	small-medium	small	Asian	Limited Liability Company(LLC)	small
1	0	U	1	medium-large	medium	medium-large	American Indian or Alaska Native	Subchapter S Corporation	medium-large
0	0	U	1	medium-large	medium-large	medium-large	White	Corporation	medium-large
0	0	U	0	small	small	small	White	Corporation	small
1	1	R	1	small-medium	medium-large	small-medium	White	Limited Liability Company(LLC)	small-medium

4.1.1 Initial Model

For our initial logistic regression model, we used data features with community and loan metrics: *cb*, *stchrtr*, *ruralurbanindicator*, *lmiindicator*, *currentapprovalamount*, *jobsreported*, *forgivenessamount*, *race*, *businesstype*, *payroll_proceed*, and *minority*. Table 5 shows the initial logistic regression's results summary.

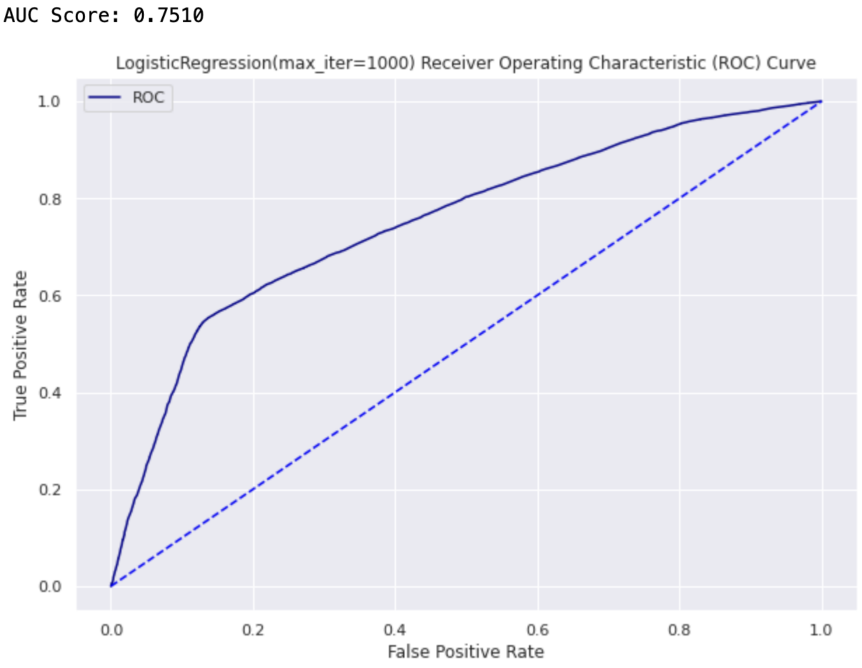
Table 5: Initial Logistic Regression Performance Metrics



```
                Logistic Regression Initial Results
-----------------------------------------------------------
***************** Evaluation on Test Data *****************

Accuracy Score 0.6864409092149426

              precision    recall  f1-score   support

           0       0.60      0.72      0.66     15243
           1       0.77      0.66      0.71     21404

    accuracy                           0.69     36647
   macro avg       0.69      0.69      0.68     36647
weighted avg       0.70      0.69      0.69     36647

-----------------------------------------------------------
```

We can observe that the model performs relatively well in precision at predicting community banks (77%) and weaker at predicting non-community banks (60%). Moreso, with recall, we can see that the model appears to switch and predict correctly at non-community banks at 72% accuracy and community banks at 66% accuracy. Recall describes the true positive predictions out of all predicted results, versus precision, which describes the proportion of true positive to actual results. The total accuracy score results are approximately 0.6864, meaning that the initial logistic regression model was able to predict the classification of community banks and non-community banks at a 68.64% accuracy. In addition to these results, Figure 5 shows the ROC (Receiver Operating Characteristic) curve of our initial model, which is a visualization of false positive rate and the true positive rate. The ROC curve has an AUC Score of approximately 0.7510, which implies that this regression is fairly accurate, especially within the true data.

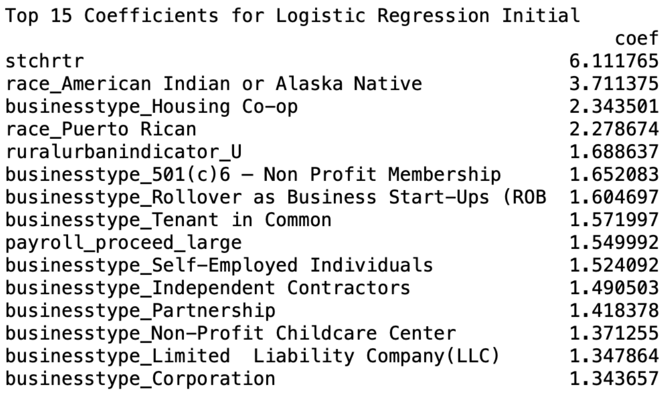
Furthermore, in order to analyze our Logistic regression model, we can examine the resulting feature coefficients to understand which features appear to be most influential when predicting if a loan is distributed by a community bank. Logistic regression coefficients represent the log odds that an observation is in the target class given the values of the various features. Table 6 below presents the Top 15 features with greatest coefficient values. For every one-unit increase in each of the top influential features, the odds that the observed loan is classified as a community is the corresponding coefficient times as large as the odds that the observation is a non-community bank when all other variables are held constant. We can notice that amongst the top influential features in predicting community banks' role are: *stchrtr*, *race_American Indian or Alaska Native, businesstype_Housing Co-op , race_Puerto Rican, ruralurbanindicator_U,* and *businesstype_501(c)6 – Non Profit Membership*.



```
AUC Score: 0.7510
```

Figure 5: Initial Logistic Regression ROC

Table 6: Top 15 Coefficients for Initial Logistic Regression

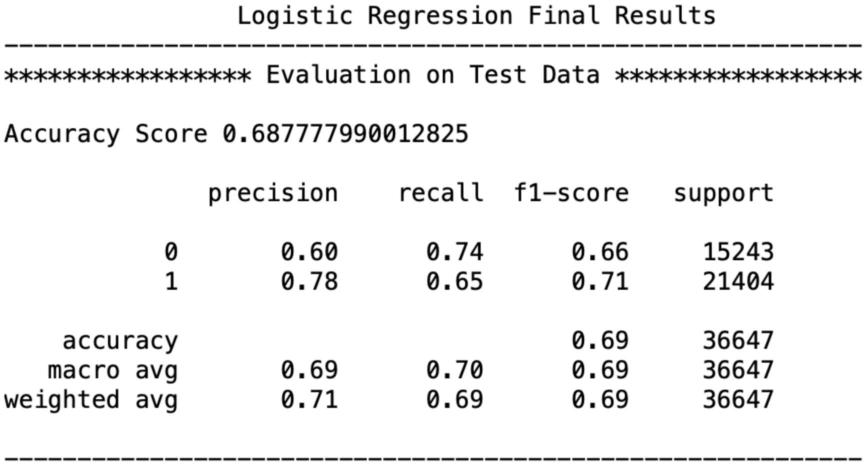


```
Top 15 Coefficients for Logistic Regression Initial
                                                      coef
stchrtr                                           6.111765
race_American Indian or Alaska Native             3.711375
businesstype_Housing Co-op                        2.343501
race_Puerto Rican                                 2.278674
ruralurbanindicator_U                             1.688637
businesstype_501(c)6 – Non Profit Membership      1.652083
businesstype_Rollover as Business Start-Ups (ROB  1.604697
businesstype_Tenant in Common                     1.571997
payroll_proceed_large                             1.549992
businesstype_Self-Employed Individuals            1.524092
businesstype_Independent Contractors              1.490503
businesstype_Partnership                          1.418378
businesstype_Non-Profit Childcare Center          1.371255
businesstype_Limited  Liability Company(LLC)      1.347864
businesstype_Corporation                          1.343657
```

4.1.2 Final Model

To further refine our Logistic regression model, we decided to remove highly correlated variables such as *forgivenessamount* and *jobsreported* that are highly correlated with *currentapprovalamount*. Our final model used the following features: *cb*, *stchrtr*, *ruralurbanindicator*, *lmiindicator*, *currentapprovalamount*, *race*, *businesstype*, and *payroll_proceed*.

Table 7: Final Logistic Regression Performance Metrics



```
                 Logistic Regression Final Results
-----------------------------------------------------------
**************** Evaluation on Test Data ****************

Accuracy Score 0.687777990012825

              precision    recall  f1-score   support

           0       0.60      0.74      0.66     15243
           1       0.78      0.65      0.71     21404

    accuracy                           0.69     36647
   macro avg       0.69      0.70      0.69     36647
weighted avg       0.71      0.69      0.69     36647

-----------------------------------------------------------
```

Looking at the final results in Table 7 for our logistic regression model, we can see that the model continues to perform relatively well in precision at an increase in accuracy at predicting community banks (78%) and continues to be weaker at predicting non-community banks (60%). For recall, we can observe that the model appears to predict correctly at non-community banks at increased 74% accuracy and community banks at 65% accuracy. This is consistent with our initial model. The total accuracy score results increased marginally at about 0.0014 for a total of approximately 0.6878, meaning that the initial logistic regression model was able to predict the classification of community banks and non-community banks at a 68.78% accuracy. Although this is a small increase in accuracy, we believe this is an improved logistic regression that is able to predict community banks with the absence of highly correlated variables. The confusion matrix as seen in Figure 6 aptly visualizes the final model's distribution of actual versus predicted weights. Our final logistic model predicts true community banks best.

Similar to our initial results, Figure 7 shows the ROC curve of our final model. The ROC curve has an AUC Score of approximately 0.7511, which implies that this classification increased by 0.0001. The ROC curve affirms that the final model is fairly accurate amongst true data.

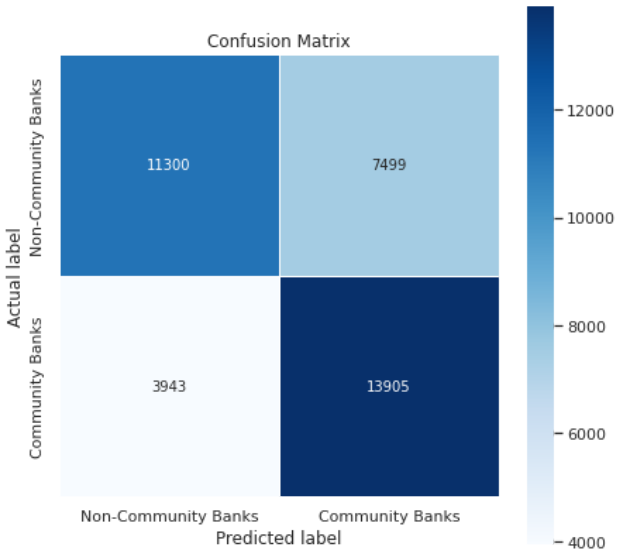


Figure 6: Confusion Matrix for Final Logistic Regression

```
AUC Score: 0.7511
```

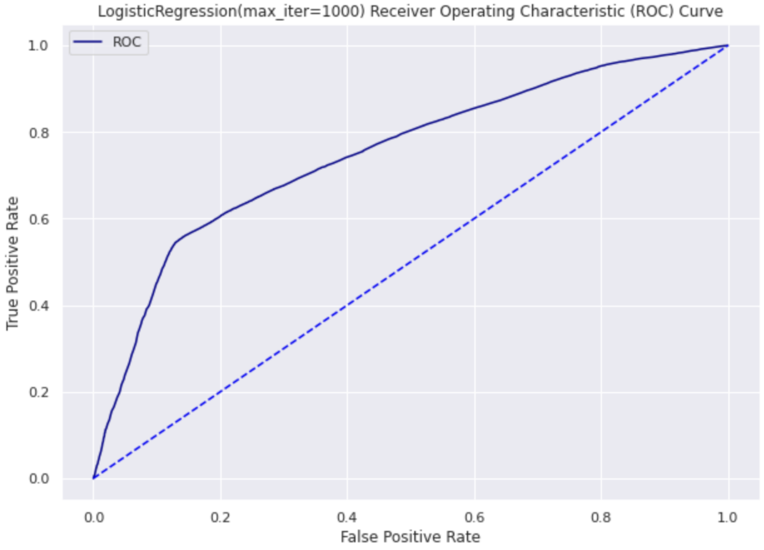
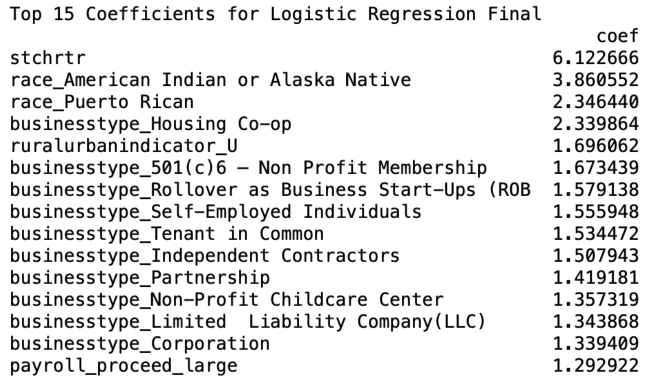


Figure 7: Final Logistic Regression ROC Curve

By refining the initial Logistic regression model, the resulting feature coefficients to understand which features appear to be most influential when predicting if a loan is distributed by a community bank altered. Table 8 below shows the Top 15 features with greatest coefficient values for the final model. It is worth noting that the top influential features in predicting community banks' role are: *stchrtr*, *race_American Indian or Alaska Native, race_Puerto Rican*, *businesstype_Housing Co-op ,* and *ruralurbanindicator_U.* Our final model shows a higher influence of Puerto Rican communities and urban areas with a decreased influence on large payroll loan allocations.

Table 8: Top 15 Coefficients for Final Logistic Regression



```
Top 15 Coefficients for Logistic Regression Final
                                                      coef
stchrtr                                           6.122666
race_American Indian or Alaska Native             3.860552
race_Puerto Rican                                 2.346440
businesstype_Housing Co-op                        2.339864
ruralurbanindicator_U                             1.696062
businesstype_501(c)6 – Non Profit Membership      1.673439
businesstype_Rollover as Business Start-Ups (ROB  1.579138
businesstype_Self-Employed Individuals            1.555948
businesstype_Tenant in Common                     1.534472
businesstype_Independent Contractors              1.507943
businesstype_Partnership                          1.419181
businesstype_Non-Profit Childcare Center          1.357319
businesstype_Limited  Liability Company(LLC)      1.343868
businesstype_Corporation                          1.339409
payroll_proceed_large                             1.292922
```

4.2 Classification and Regression Trees

With our second model, a Classification and Regression Tree model, we attempted to predict a class and numeric label for community and non-community bank loans. To find the best split in the tree, we calculated the weighted sum of Gini Impurity for both child nodes. In our fitting function, we continued doing this for all possible splits then took the one with the lowest Gini Impurity as the best split.

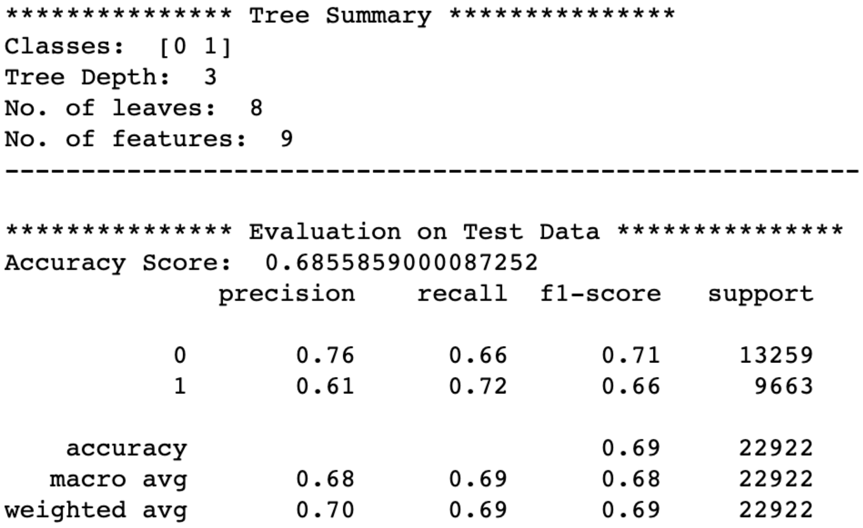
$$Gini\ Impurity = 1 - Gini = 1 - \sum_{i=1}^{n} p_i^2$$

The fitting function was also used to split the data into training (30%) and testing (70%) samples. The model was then fit with DecisionTreeClassifier with the parameters of criterion, splitter, max_depth, class_weight, min_samples-leaf, and random_state. The criterion, or the function that measures the quality of a split, was set to "Gini" to utilize the Gini impurity for the information gain. The strategy to choose the split at each node was set to its default to choose the best split and the maximum depth was set to 3 with the minimum number of samples required to be at a leaf node set 1000. By limiting tree depth and leaf size, we are able to avoid some initial overfitting.

4.2.1 Initial Model

In our initial model, we considered the variables *stchrtr, ruralurbanindicator, payroll_proceed, minority, jobsreported, forgivenessamount,* and *currentapprovalamount* to predict if a loan originated from a community bank.

Table 9: Initial CART Model Performance Metrics



```
*************** Tree Summary ***************
Classes:  [0 1]
Tree Depth:  3
No. of leaves:  8
No. of features:  9
--------------------------------------------------------

*************** Evaluation on Test Data ***************
Accuracy Score:  0.6855859000087252
              precision    recall  f1-score   support

           0       0.76      0.66      0.71     13259
           1       0.61      0.72      0.66      9663

    accuracy                           0.69     22922
   macro avg       0.68      0.69      0.68     22922
weighted avg       0.70      0.69      0.69     22922
```

Initial model results yielded 68.5% accuracy on testing data as seen in Table 9. We can see that the model performs relatively well in precision at predicting non-community banks (76%) and slightly lower at predicting community banks (61%). Recall, however, appears to be better with community banks at 72% and non-community banks at 66% accuracy. To note, precision means that the loan is actually from a community bank in 61% of cases and recall means that for all of the community bank loans in the test data, the model identified 72% of them. The difference between precision and recall for community and non-community bank loans could be attributed to an imbalance in the data, which we explored more later.

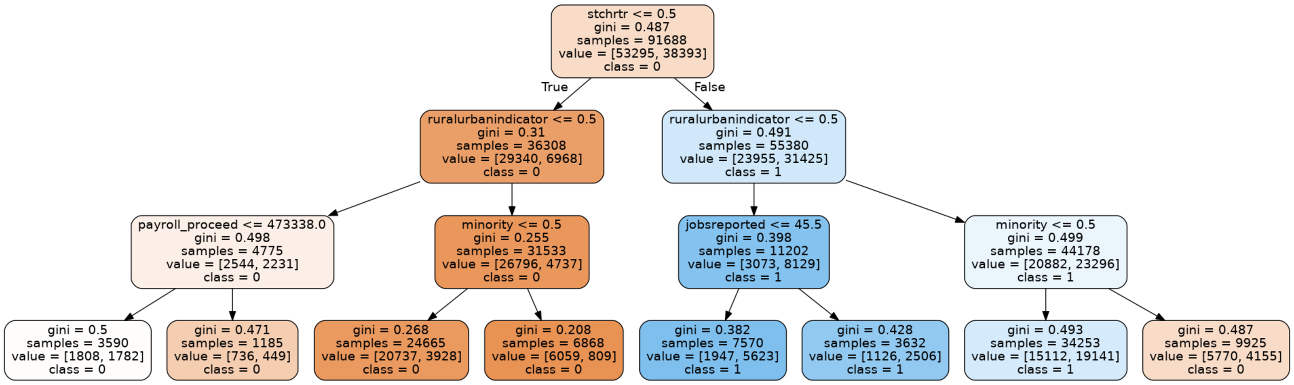


Figure 8: CART Initial Model Decision Tree

Considering the tree graph generated by the initial model in Figure 8, we can see that the fitting function used several different variables to generate splitting criterion. It appears that the two most important variables from the initial model for determining community banks were *stchrtr*, *ruralurbanindicator*, and *minority* since they influenced a class label prediction of 0 or 1.

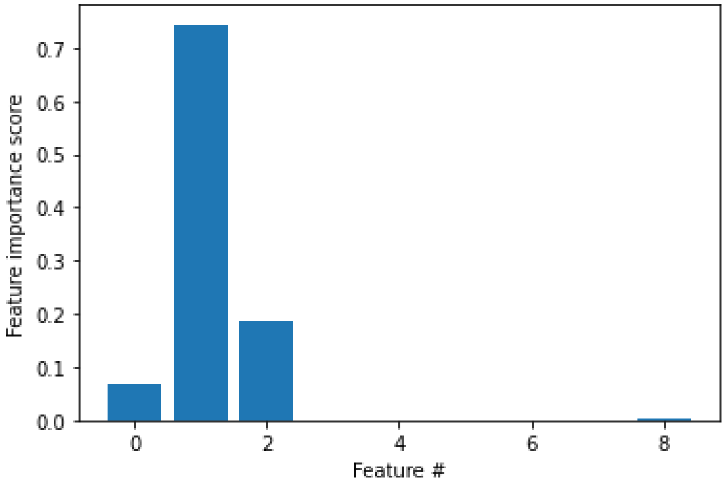


Figure 9: CART Initial Model Feature Importance

Table 10: Initial CART Detailed Feature Importance

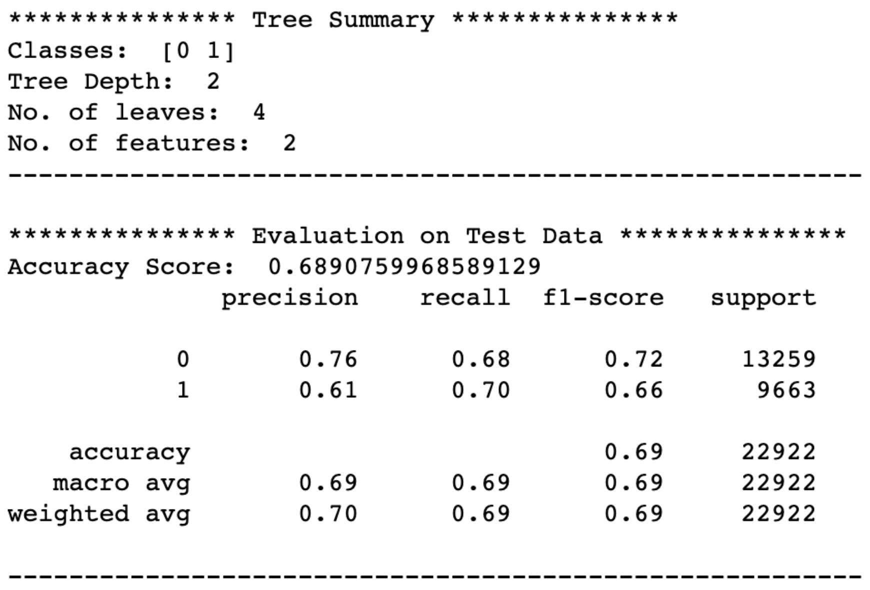
Feature Number	*Variables*	*Feature Importance Score*
0	minority	0.06912
1	stchrtr	0.74371
2	ruralurbanindicator	0.18486
3	jobsreported	~0
4	currentapprovalamount	~0
5	forgivenessamount	~0
6	lmiindicator	~0
7	businesstype	~0
8	payroll_proceed	0.00231

Using the CART algorithm for feature importance implemented in the *scikit-learn* as *DecisionTreeRegressor* and *DecisionTreeClassifier* classes, we can fit the model to analyze feature importance based on the same parameters. After being fit, we used the model's *feature_importances_property* to calculate importance scores for each input variable as seen in Figure 9 and Table 10. With these scores, we can validate the results displayed in Figure 8 for feature importance with *minority*, *stchrtr,* and *ruralurbanindicator* ranking the highest.

4.2.2 Reduced Model

After experimenting with variable combinations based on the initial model's feature importance and tree graph, we selected the variables *stchrtr* and *minority* to create a reduced model.

Table 11: CART Reduced Model Performance Metrics



```
*************** Tree Summary ***************
Classes:  [0 1]
Tree Depth:  2
No. of leaves:  4
No. of features:  2
-------------------------------------------------------

*************** Evaluation on Test Data ***************
Accuracy Score:  0.6890759968589129
              precision    recall  f1-score   support

           0       0.76      0.68      0.72     13259
           1       0.61      0.70      0.66      9663

    accuracy                           0.69     22922
   macro avg       0.69      0.69      0.69     22922
weighted avg       0.70      0.69      0.69     22922

-------------------------------------------------------
```

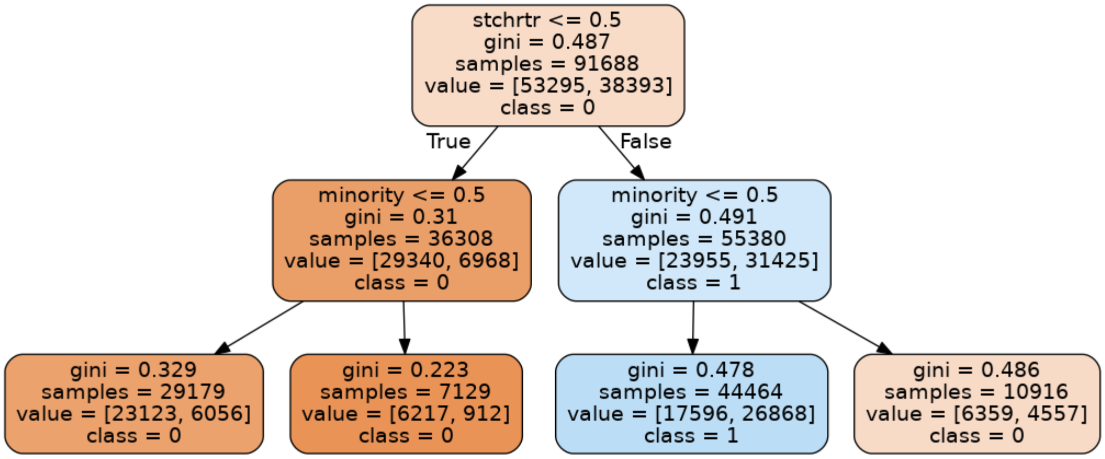


Figure 10: CART Reduced Model Tree Graph

The reduced model results yielded 68.9% accuracy on testing data as seen in Table 11 (+0.5% from the initial model). We can see that the model performs relatively well in precision at predicting non-community banks (76%) and slightly lower at predicting community banks (61%). Recall, however, appears to be better with community banks at 70% and non-community banks at 68% accuracy. The reduced model's accuracy may not have changed significantly from the initial, however, it demonstrates the predicting power of the variables of importance identified by the initial model. By removing extraneous variables and looking at *stchrtr, minority,* and *ruralurbanindicator* alone, we can generate similar accuracy at predicting community banks as considering the nine original variables together.

4.3 Naive Bayes

Our final model is Naive Bayes which makes use of probabilities, a 'most-likely' method, to predict whether a bank is a community vs non-community bank. The Naive Bayes algorithm takes in the given variables and then returns a probability score for the variable being predicted, community bank vs non-community bank. In this algorithm, each feature used to give a prediction is assumed to be independent, meaning that one variable is not related to another. In order to do this, the Naive Bayes algorithm utilizes Bayes' law as seen below:

$$P(C \mid A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A \mid C)P(C)}{P(A)}$$

In our model C would represent the variable *cb* (community vs non-community bank) while A represents an observation with the features being used to predict in the corresponding model to produce the formula below:

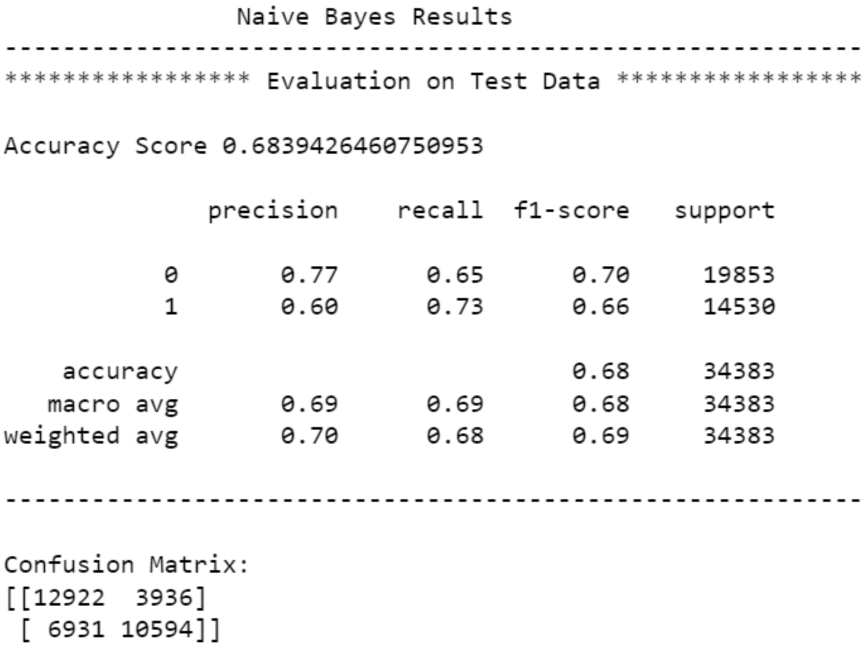
$$P(cb \mid features) = \frac{P(features \mid cb)P(cb)}{P(features)}$$

In order to use this model we must convert every variable into a probability or frequency. This model assumes that all features have an equal weight, eliminating any possible human biases in determining importance of a variable. Since all variables are treated ‘equally’ within this model, it is important to remove features that are highly correlated/related because it will result in the model counting it ‘twice’. Examples of this would be race and ethnicity,

4.3.1 Initial Model

For our initial Naive Bayes model, we used *borrowerstate, ruralurbanindicator, race, businesstype, stchrtr, lmiindicator, currentapprovalamount, forgivenessamount, minority,* and *payroll_proceed* to determine *cb* (community bank vs non-community bank).

Table 12: Initial Naive Bayes Model Performance Metrics



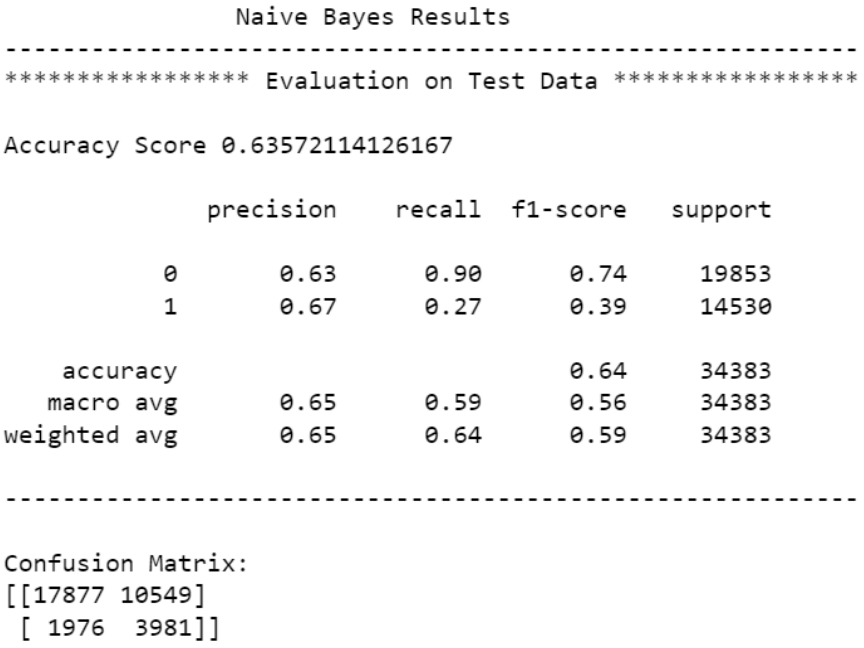
```
               Naive Bayes Results
-----------------------------------------------------
**************** Evaluation on Test Data ****************

Accuracy Score 0.6839426460750953

              precision    recall  f1-score   support

           0       0.77      0.65      0.70     19853
           1       0.60      0.73      0.66     14530

    accuracy                           0.68     34383
   macro avg       0.69      0.69      0.68     34383
weighted avg       0.70      0.68      0.69     34383

-----------------------------------------------------

Confusion Matrix:
[[12922  3936]
 [ 6931 10594]]
```

This initial model has an accuracy of 68% which is on par with the previous models described. We can see that the model performs relatively well in precision at predicting non-community banks (77%) and slightly lower at predicting community banks (60%). Recall, however, appears to be better with community banks at 73% and non-community banks at 65% accuracy, similar to the previous CART model.

4.3.2 Final Model

For our final model we used *ruralurbanindicator, race, businesstype, stchrtr, lmiindicator, currentapprovalamount,* and *payroll_proceed* to predict *cb* (community bank vs non-community bank).

Table 13: Final Naive Bayes Model Performance Metrics



```
                Naive Bayes Results
-----------------------------------------------------------
**************** Evaluation on Test Data ****************

Accuracy Score 0.63572114126167

              precision    recall  f1-score   support

           0       0.63      0.90      0.74     19853
           1       0.67      0.27      0.39     14530

    accuracy                           0.64     34383
   macro avg       0.65      0.59      0.56     34383
weighted avg       0.65      0.64      0.59     34383

-----------------------------------------------------------

Confusion Matrix:
[[17877 10549]
 [ 1976  3981]]
```

Although the accuracy of this model is lower than the original by 4% it is important to note that the only variable that reduced the accuracy from the first to the second model was *forgivenessamount.* Removing the other variables only affected the model marginally. Although removing the *forgivenessamount* variable reduced the accuracy, the variable was removed from the model because it is highly correlated with *currentapprovalamount*. Since the majority of PPP loans have been forgiven, the two variables are highly similar and are being ‘double-counted’ within the model which we want to avoid when using Naive Bayes.

5.0 Results and Analysis

Through our data analysis and modeling we were able to find the answers to our questions of whether there were statistically significant differences between business size, type, and ownership for the businesses that received PPP loans from community banks vs non-community banks.

Only businesses with 500 people or fewer were eligible to apply for a PPP loan so all businesses from our dataset had to be that size or smaller. When looking at our data the *jobsreported* variable is the variable we used to indicate the business size. When exploring the data and looking at correlations, we discovered that this variable was not as significant as we originally expected so it was eliminated when producing final prediction models.

Unlike business size, business type was discovered to be more significant amongst the different variables in our dataset. Due to this, the variable *businesstype* was used when producing our prediction models. Although our final CART model did not use this variable to conduct predictions, both the Logistic Regression and Naive Bayes models made use of the variable to make predictions.

Lastly, when looking into business ownership we took into consideration the following variables: *race, minority, ethnicity, veteran, gender, etc.* Additionally we looked at variables related to ownership such as *lmiindicator, ruralurbanindicator,* and *stchrtr*. When analyzing and conducting statistical tests on these variables we discovered that some variables such as *veteran* and *gender* were not as significant as we anticipated, likely due to the large number of null values amongst the observations. *Race, minority,* and *ethnicity* are variables very similar in definition and once we created our models, we discovered that only one of each of the variables needed to be included to make predictions. Both the Logistic Regression and Naive Bayes models made use of the *race* variable to make predictions while CART utilized the *minority* variable to make its predictions. Additionally, all three finalized models made use of the *stchrtr* variable while only Logistic Regression and naive bayes used *lmiindicator* and *ruralurbanindicator* to make predictions. The following table compares the three models:

Table 14: Model Performance Metrics Comparison

Models:	Logistic Regression	CART	Naive Bayes
Variables Utilized			
stchrtr	X	X	X
ruralurbanindicator	X		X
lmiindicator	X		X
currentapprovalamount	X		X
race	X		X
businesstype	X		X
payroll_proceed	X		X
minority		X	
Model Details			
Accuracy	69%	69%	64%
Non-Community Bank Precision	60%	76%	63%
Community Bank Precision	78%	61%	67%

From the table we can see that both Logistic Regression and Naive Bayes utilized the exact same variables to make their predictions. The Logistic Regression and CART models had the highest accuracies at 69% but one model did better at predicting non-community banks vs community banks. The Logistic Regression model had a higher precision for community banks at 78% while the CART Model had higher precision for non-community banks at 76%. The results of the Naive Bayes model were very similar to that of the Logistic Regression model, just with lower percentage points.

6.0 Conclusion and Next Steps

The Paycheck Protection Program (PPP) was developed to help struggling small businesses during the COVID-19 pandemic. The loans were intended to be used to assist small businesses keep their workforce employed since many businesses were struggling and letting workers go during this time period. After the first round of PPP loans were distributed to small businesses there was much chatter about how the distribution of the loans were inequitable and unfair to particular small businesses. While it was found that normally disadvantaged groups were less likely to have received the PPP loans, we wanted to find out if community banks had a different impact on the distribution of these loans.

Our original hypothesis was that there are statistically significant differences in the business type, size, and ownership of the businesses that received PPP loans from community banks vs non-community banks. From our analysis we discovered that only business type and some ownership factors differentiated loans distributed from community banks and non-community banks. Size was not a factor in this situation. Additionally, we discovered that some factors like *payroll_proceed* were also significant in determining whether the loan was given out by a community bank or non-community bank. This factor makes a lot of sense when thinking about community banks and how closely connected they are to their customers as opposed to non-community banks. For instance, one possible explanation could be that community banks want PPP dollars to go towards the workers as opposed to utility or rent bills of the small business.

Looking forward, it would be interesting to see how the distribution of the second round of PPP loans differentiated from the first round that we looked at for this project. Did the backlash and headlines affect the distribution of the second round of loans or did it remain the same? Are the significant variables in community bank classification constant or did they change with the second round? While it makes sense that not all businesses that applied for a PPP loan received one, we would also be curious to investigate the characteristics of businesses that did not receive loans. Would a business that was rejected for a PPP loan through a non-community bank have been able to receive a PPP loan from a community bank?

Policymakers should consider these questions in future legislation and policy implementation. In an April SBA press release, it was announced that small business standards were revised to increase eligibility for federal loan programs. Given our analysis, we hope these standards will continue to push in the direction of the variables we identified to expand access to much needed financial aid for small businesses. We would also encourage SBA to include minority, low-income, and rural identifiers in the loan process to ensure at-risk populations are not forgotten. We recognize fraud as a potentially legitimate risk to the stability of community banks and other financial institutions and support SBA's recent announcement that SBA Administrator Isabella Casillas Guzman has publicized efforts to crack down on fraud in loan programs. SBA should consider expanding loan education programs and corporate partnerships to support businesses specifically serviced by community banks as an effort to bolster resilience against crises and increase awareness around loan opportunities. With the adoption of these policies, we hope to see our findings utilized as a next step in advancing the mission of CSBS- to ensure safety and soundness, promote economic growth, and enable historically underserved and disadvantaged communities.

7.0 References

Administrator Guzman announces expanded efforts to aggressively crack down on bad actors and prevent fraud in programs. (n.d.). Retrieved April 24, 2022, from https://www.sba.gov/article/2022/apr/01/administrator-guzman-announces-expanded-efforts-aggressively-crack-down-bad-actors-prevent-fraud

Autor, D., Cho, D., Crane, L. D., Goldar, M., Lutz, B., Montes, J. K., Peterman, W. B., Ratner, D. D., Vallenas, D. V., & Yildirmaz, A. (2022, January 17). *The $800 billion paycheck protection program: Where did the money go and why did it go there?* NBER. Retrieved April 24, 2022, from https://www.nber.org/papers/w29669

Beginners Guide to naive Bayes algorithm in Python. Analytics Vidhya. (2021, January 16). Retrieved April 24, 2022, from https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/

Brownlee, J. (2020, August 20). *How to choose a feature selection method for machine learning*. Machine Learning Mastery. Retrieved April 24, 2022, from https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brownlee, J. (2020, December 9). *Information gain and mutual information for Machine Learning*. Machine Learning Mastery. Retrieved April 24, 2022, from https://machinelearningmastery.com/information-gain-and-mutual-information/

IBM Cloud Education. (n.d.). *What is supervised learning?* IBM. Retrieved April 24, 2022, from https://www.ibm.com/cloud/learn/supervised-learning

Centers for Disease Control and Prevention. (n.d.). *Basics of covid-19*. Centers for Disease Control and Prevention. Retrieved April 24, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-19.html

Community banks: Number by state and asset size. Banking Strategist. (n.d.). Retrieved April 24, 2022, from https://www.bankingstrategist.com/community-banks-number-by-state-and-asset-size#:~:text=ARE%20LOCAL%20BANKS.-,Community%20Banks%20are%20Local%20Banks.,higher%20concentration%20of%20community%20banks

CSBS 2022 annual data analytics competition: The role of Community Banks during the pandemic. CSBS. (n.d.). Retrieved April 24, 2022, from https://www.csbs.org/csbs-2022-annual-data-analytics-competition-role-community-banks-during-pandemic

D;, S. N. S. (n.d.). *Supervised mutual-information based feature selection for Motor Unit Action Potential Classification*. Medical & biological engineering & computing. Retrieved April 24, 2022, from https://pubmed.ncbi.nlm.nih.gov/9538543/

Did asian- and American Indian-owned businesses receive their fair share of PPP loans? Heartland Forward. (2021, March 4). Retrieved April 24, 2022, from https://heartlandforward.org/case-study/did-asian-and-american-indian-owned-businesses-receive-their-fair-share-of-ppp-loans/

Jain, A. (2020, July 19). *Learn how to do feature selection the right way*. Medium. Retrieved April 24, 2022, from https://towardsdatascience.com/learn-how-to-do-feature-selection-the-right-way-61bca8557bef

Paycheck protection program. (n.d.). Retrieved April 24, 2022, from https://www.sba.gov/funding-programs/loans/covid-19-relief-options/paycheck-protection-program

Paycheck protection program. U.S. Department of the Treasury. (2022, January 12). Retrieved April 24, 2022, from https://home.treasury.gov/policy-issues/coronavirus/assistance-for-small-businesses/paycheck-protection-program

PPP FOIA - U.S. Small Business Administration (SBA): Open data. U.S. Small Business Administration (SBA) | Open Data. (2022, April 5). Retrieved April 24, 2022, from https://data.sba.gov/dataset/ppp-foia

Saxena, S. (2018, May 13). *Precision vs recall*. Medium. Retrieved April 24, 2022, from https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488

SBA Revises Small Business Size Standards. SBA revises Small Business Size Standard in 16 industrial sectors to increase eligibility for its federal contracting and Loan Programs. (n.d.). Retrieved April 24, 2022, from https://www.sba.gov/article/2022/apr/04/sba-revises-small-business-size-standards-16-industrial-sectors-increase-eligibility-its-federal

Small Business Digital Alliance publishes library of free digital tools from national members, fortune 500 companies available to small businesses. (n.d.). Retrieved April 24, 2022, from https://www.sba.gov/article/2022/mar/31/small-business-digital-alliance-publishes-library-free-digital-tools-national-members-fortune-500