Raymond A. Mason
School of Business
WILLIAM & MARY

Quantifying the Role Community Banks' Fintech Partnership Played in PPP Loan Distribution

CSBS 2022 Annual Data Analytics Competition Final Report

Competition Members:

Gio DeFrank

Junghee Mun

Kristina Posner

Faculty Advisor:

Professor Joseph Wilck

I. Executive Summary:

The 2022 Conference of State Bank Supervisors' (CSBS) Annual Data Analytics Competition assigned students the task of analyzing and understanding the role community banks played within the Covid-19 pandemic. With the number of community banks falling due to financial pressures, it is essential we understand how impactful community banks can be to our society especially in economic downfalls. CSBS provided students with a database of Paycheck Protection Program (PPP) loans during the Covid-19 pandemic to quantify the impact that less than 5,000 U.S. community banks played in our economy. Initially, teams were given the task of developing a hypothesis about the role or impact community banks played during the pandemic. Since the banking industry is forever evolving, it is important to also understand how community banks utilize technology to help society. Therefore, our group's hypothesis focuses on if community banks who partnered with financial technology (fintech) firms were ultimately more successful in distributing PPP loans to small businesses with racial minorities and underbanked populations.

To test our hypothesis, we utilized two dependent variables: minority (binary variable indicating if borrower was a minority) and LMI (binary variable indicating if borrower lived in low and moderate income communities) to test how fintech partnership impacted the amount of loans given to minorities and those in low income communities. We utilized a random forest model to predict minority, achieving an accuracy rate of 66.83%. We followed this approach once again to predict LMI, achieving an accuracy rate of 69.864%. Each model's importance report showed that fintech partnership had the littlest impact when predicting if the borrower was a minority or lived in a LMI. Therefore, a community bank's reach to minorities or those in lower income areas was not drastically different if they partnered with fintech companies. If a

community bank's main motivation in partnering with fintech companies is to increase their outreach to minorities and LMI, they will need to better their approaches used during Covid-19.

II. Related Literature and Research:

A. Current Literature Overview

Research is just starting to develop on the impact the pandemic had on different players within the banking industry. Preliminary research by several organizations (including the FDIC and CSBS) shows that community banks played an outsized role in the distribution of PPP loans to small businesses. Likewise, recent studies done by the National Bureau of Economic Research examined the impact fintechs had during the pandemic – specifically focusing on the effect fintechs had on distributing loans to minority-owned businesses. However, while many anecdotal reports and stories discuss the impact community banks partnered with fintechs had on the successful distribution of PPP loans, no formal research appears to have been done on the subject.

B. Model Summary and Justification

To test the results of a fintech partnership on community bank's outreach to minority-owned businesses and lower-income communities, we implemented several models. Initially, our group implemented a binary logistic regression to test the impact of fintech partnership on the odds of a PPP borrower being a minority. However, the logistic regression model must adhere to four assumptions for the model results to be seen as reliable. Since our established model violated half of the assumptions, we implemented a random forest model using the ranger package. A classification random forest model will run a series of decision trees, eventually taking the most predicted classification values (whether an individual is a minority or

not) to establish the model's overall predicted value. It provides a lot more flexibility compared to the logistic regression and reduces the risk of overfitting. When we ran a random forest model on the minority variable, our model achieved a 66.83% accuracy rate. We then proceeded to run a random forest model on the LMI variable, achieving a 69.864% accuracy rate. Therefore, we could confidentially interpret each model's importance report to determine how impactful fintech partnership was when determining our dependent variable.

III. Data Sources:

A. PPP Loan Data

As part of the competition, PPP loan data containing over eight million rows with columns ranging from the specific loan number to the minority status of the business requesting a PPP loan, was provided. This data formed the core of our analysis, as it provided loan-level information by which to test our hypothesis.

B. FDIC Community Bank Designation Data

In addition to the PPP loan-level data provided, we also decided to include several other data sources to more holistically understand our business question and subsequent insights. For example, we connected the PPP loan data with data from the FDIC which detailed the specifications required to be officially labeled as a community bank. This dataset not only allowed us to segment our PPP loan data by community and non-community banks according to the FDIC's official definition, but it also provided a wealth of information relating to each bank's financials – including assets, loan-to-asset ratios, office (branch) counts, and much more.

C. *FDIC Yearly How America Banks Survey*

Finally, data relating to a state's unbanked population was also combined with the PPP loan data to provide insight into how fintech and community bank partnerships may have impacted areas with high unbanked populations. This data was pulled from the FDIC's yearly national "How America Banks" survey database for the years 2015-2019 (2020 wasn't yet available). This dataset included the FDIC's estimated proportion of each state's population that had or didn't have a bank account, as well as the confidence intervals for each of these categories.

D. *Aggregated community bank information*

The CSV file is segmented by bank total assets size in order to present trends across all segments of Community Banks (banks and thrifts with under $10 billion in total assets). This view provides a better perspective on activity and changes among the large population of community banks and thrifts that serve local communities across the U.S. and where these community banks and thrifts are chartered and located. Data is as of Q4 2021.

IV. Data Collection and Cleaning Methodology:

After deciding on the data sources required to successfully test the hypothesis, each one was examined to determine the level of processing and cleaning required to extract useful insights.

A. Web scraping

As our hypothesis revolves around the interaction between community banks and fintechs – and the part this interaction played in the success of PPP loan distribution – a list of community banks that partnered with fintechs during the pandemic was required. However, as the subject is

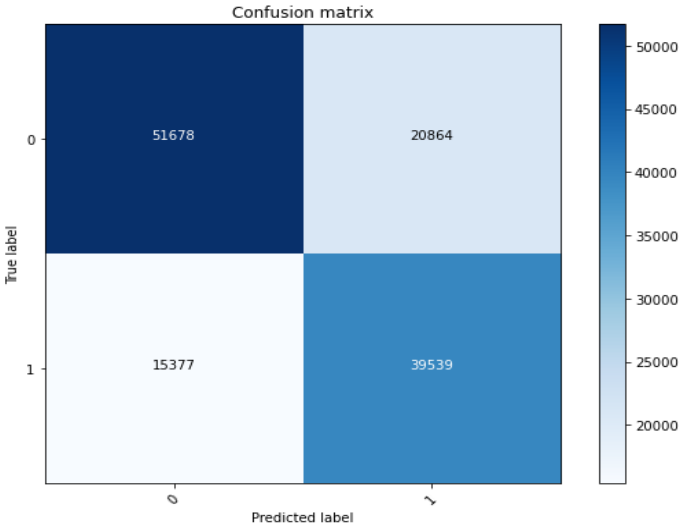
still relatively new, and the hypothesis untested by others, no comprehensive list already existed from which we could draw. Realizing this, we created our own list by web scraping google for links that mentioned community banks and fintech partnerships in the distribution of PPP loans. We generated over 100 links and systematically examined each one for reference to a specific community bank and fintech company. We recorded each partnership, validated it against the bank and fintech websites, and verified the institution's status as a community bank with the FDIC. Once this list was complete, a total of 41 partnerships were identified. This information was combined with our current data sources, allowing us to classify about 237,000 PPP loans that were likely distributed by community bank and fintech partnerships.

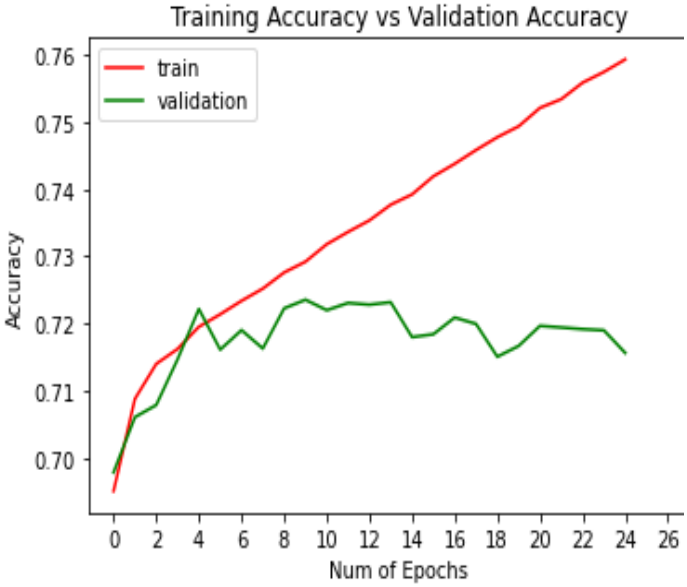
B. Classification Neural Network Model

The next challenge present in our dataset was the existence of many null values in key columns. Specifically, our hypothesis wanted to test the importance of a community bank fintech partnership on the distribution of PPP loans to minority-owned businesses. However, about 80% of the minority column was null, which if left unaltered, would severely reduce the number of samples that could be used to evaluate our hypothesis. To remedy this issue, a portion of the 20% of samples that contained a value in the minority column was extracted from the dataset and used to train a artificial neural network (ANN) for classification. This ANN model was designed to predict the likelihood of a business being minority-owned. The model followed a feedforward design, using six hidden layers and well-known activation and optimization functions ('relu' and 'adam,' respectively) to train itself. After much experimentation, this model was able to accurately predict the minority status of a business about 72% of the time on new data unused in the training of the model. While we would've preferred a higher accuracy, we believe an

accuracy rate of 72% is a result of limitations within the data versus any limitations in the actual model design – as such, the model was used to fill in the minority column of our dataset.

Figure #1 & #2: AI Confusion Matrix and Accuracy by Epoch



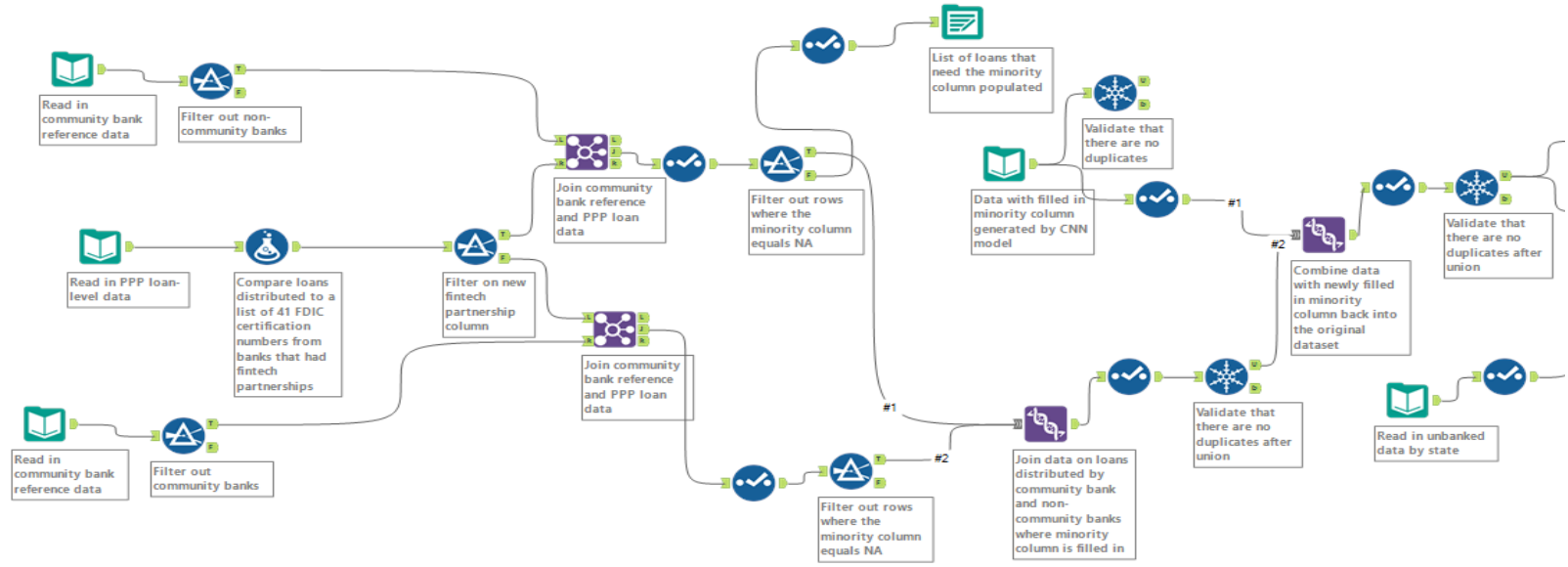


* File 6

C. Alteryx Workflow

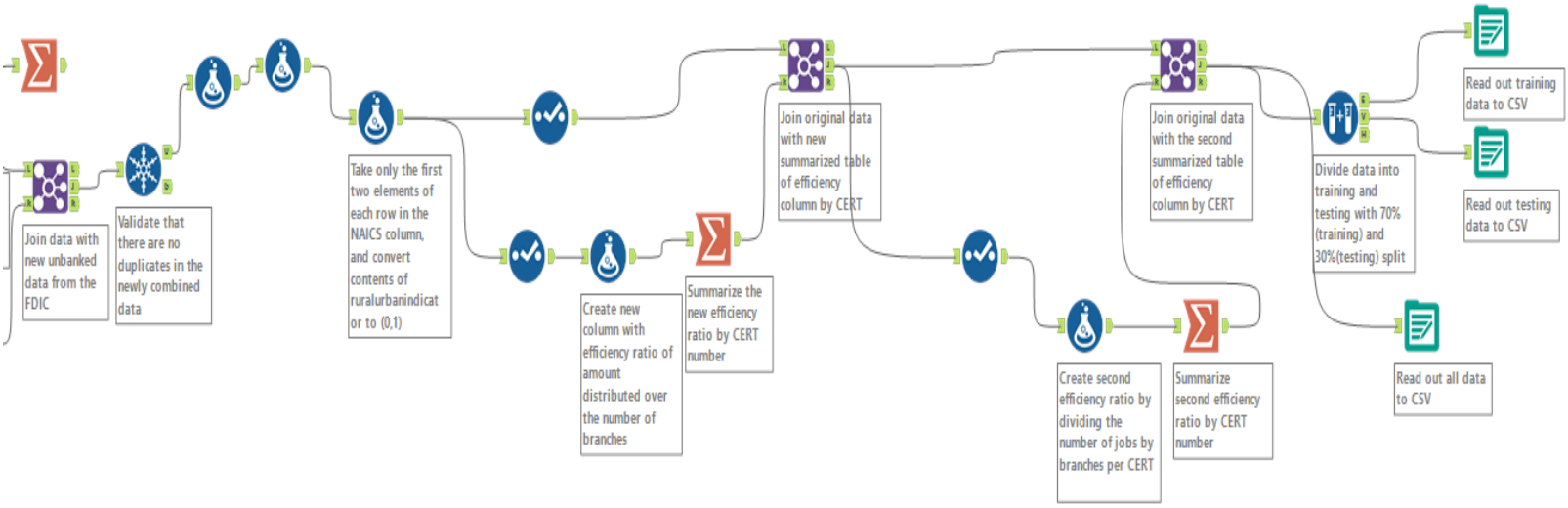
Finally, having identified our three data sources, filled in the minority column, and web scraped a list of community bank and fintech partnerships, we combined and cleaned everything using Alteryx. The platform allowed us to easily join, clean, and validate each data source. Furthermore, as the platform has many statistical and machine learning related modules, we were able to generate our test and train data sets required for our model analysis (detailed below) directly in the program.

Figure #3: Alteryx Workflow



* File 10

Figure #4: Alteryx Workflow



* File 10

V. Data Dictionary

	Variable	Explanation
1	Asset	Total assets of the institution (in thousands)
2	LoanToAsset	Loan to asset ratio of the organization
3	CoreRatio	Core deposit ratio of the organization
4	Office_Count	Number of offices in which the organization operates
5	Unique_Metros	Number of unique large metro areas in which the organization operates (has branches).
6	State_Count	Number of states in which the organization has branches.
7	ruralurbanindicator	Binary variable; 1 if urban, 0 if rural
8	lmiindicator	Binary variable; 1 if considered Low/Median Income, 0 otherwise
9	minority	Binary variable; 1 if minority, 0 otherwise
10	currentapprovalamount	PPP loan approval amount (in thousands)
11	jobsreported	Number of jobs reported per PPP loan application
12	FintechPartnership	Binary variable; 1 if partnered with Fintech for PPP loan processing, 0 otherwise
13	Number.of.Households.. 1000s.	Estimated number of households (in 1000s) within state
14	Unbanked	Percentage of unbanked population in a state

* File 14

VI. Minority Model Methodology:

A. Overview:

In May 2020, the U.S. Census Bureau released estimates showing over 1 million employer firms were owned by minorities. From a policy standpoint it is incredibly important to understand and measure how banks can best reach minority groups within our country. A big incentive for banks to partner with fintech companies is to reach more groups, specifically groups, such as minorities, who may lack the resources needed to apply for financial programs. Our team decided to set our created "minority" feature as the dependent variable to analyze what features contributed heavily to a PPP loan being given to a minority group. Since each borrower in our dataset was classified as a 0 (not minority) or 1 (minority), we ran two models: classification logistic regression and a classification random forest model to account for differences in distribution.

B. Classification Logistic Regression Minority Model:

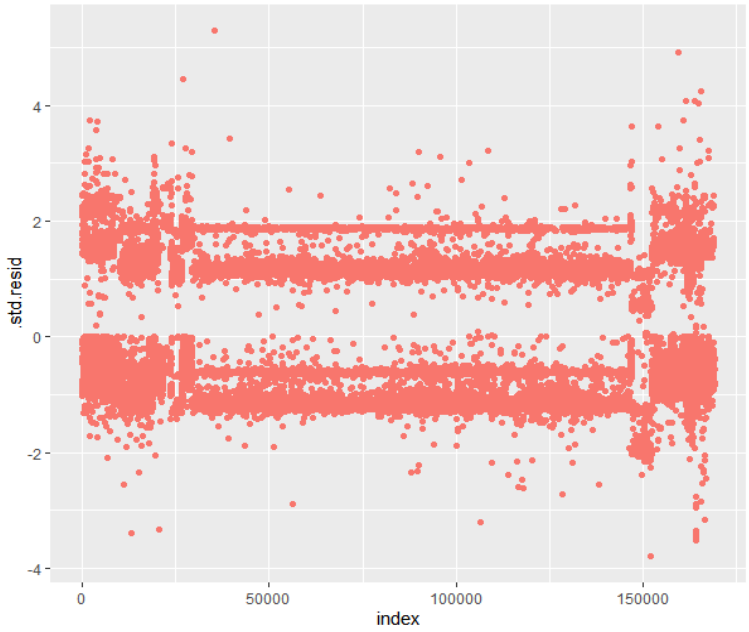
Initially, we ran a binary logistic regression model on the dependent variable "minority", achieving an accuracy level of 61.83%. However, we found our logistic regression model's results were unreliable since they violated two out of four assumptions of a logistic regression model.

We first tested the assumption of linearity of the logit to ensure there was a linear relationship between every continuous variable and their logit-transformed outcome. Therefore, our team created a statistical term representing the interaction between each of the 10 continuous variables in our analysis and their natural logarithms. Each of the 10 statistical terms were statistically significant, therefore violating this assumption.

Next, we tested the absence of multicollinearity, hoping to find an absence of multicollinearity within our independent variables. Our team ran a Variance Inflation Factor (VIF) score test, to see how well each independent variable is explained by other independent variables. We found the highest VIF score belonging to “State_Count” with a score of 3.54. Typically, any VIF score between 5-10 indicates multicollinearity, therefore our model does not violate the assumption of multicollinearity.

Then, we tested for the lack of strongly influential outliers. Therefore, we tested our model’s predicted outcomes to the actual outcomes to determine if they act as an outlier (greater than 3 standard deviations). As shown in the plot below, this assumption is violated because a large group of observations fall outside the 3 standard deviations threshold, acting as strongly influential outliers.

Figure #5: Influential Outliers



* File 14

The last assumption was independence of errors. Since there was an autocorrelation of errors, logistic regression may not be fully explaining the variance in the dataset. This is because the distribution of errors is influenced by or correlated to the errors in prior observations. Dependence of errors may arise in our dataset because each loan was processed at different times but aggregated characteristics, such as state demographics, were captured cross-sectionally. Factors surrounding each loan application may not be unique from another application, creating an overlap of errors between variables.

C. Classification Random Forest Minority Model:

In order to improve the integrity of our model results, we switched from the logistic regression model to a random forest model. Randomly forest was selected for its strong track record in classification problems, ease of interpretability, and lack of assumptions. Since our dataset was so large: test data contained 72,812 observations and train data contained 168,968 observations, we utilized the ranger package to run this model quicker.

We ran 5 different random forest models, each with a different number of trees (100, 200, 300, 400, 500) to split the data. Within each model, we ran a random forest model on "minority" using our training data, changing the number of trees in each model. Once our model ran, we proceeded to make predictions using our established model on our test data, comparing our predictions to the actual test values, creating an accuracy rate from the results. As shown in the table below, each model contained a different accuracy level. We chose the model with 300 trees as our final model since it had the highest accuracy rate (66.83%).

Figure #6: Random Forest Accuracy By Tree (Minority Prediction)

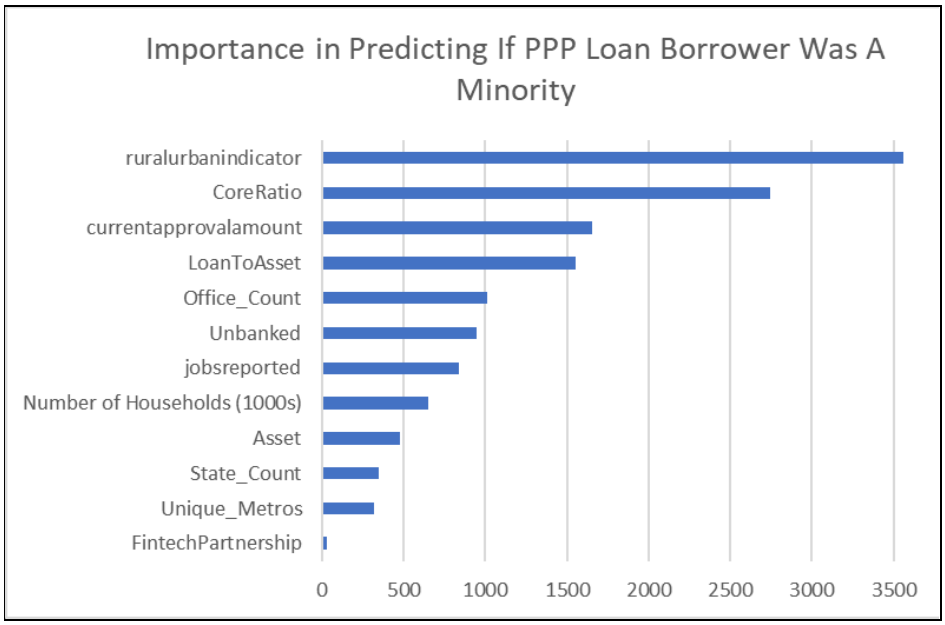
Random Forest Models	
Number of Trees	Accuracy
100	66.73%
200	66.76%
300	66.83%
400	66.77%
500	66.77%

* Derived from File 14

D. Minority Model Results:

To see how our independent variables, including fintech sponsorship impacted "minority" we printed the importance report for our final model. As shown below, the variables that had the greatest impact on whether a PPP borrower was a minority was "ruralurbanindicator" and "CoreRatio". In relation to our team's hypothesis, the variable indicating if a community bank partnered with a fintech had the least amount of importance when it came to predicting if a borrower was a minority.

Figure #7: Random Forest Variable Importance (Minority Prediction)



* Derived from File 14

E. Minority Model Importance:

From a policy perspective, these results are incredibly important. It shows that fintech partnership for community banks isn't a successful way for them to reach minority borrowers. This could be for several reasons, a community bank may have a loyal customer base already, therefore, a fintech partnership may have only reached their existing customers, which may have lacked minority groups. Another reason could be minority groups lacked the resources to take advantage of community bank fintech programs, thereby being unaware of a community bank's PPP loan distribution despite their fintech partnership. In order to increase minority community outreach through fintech partnership in future economic downfalls, community banks will need to analyze their previous tactics used during Covid-19.

VII. Low and Moderate Income Model Methodology:

A. Overview:

Our group decided to set our created "LMI Indicator" feature as the dependent variable to analyze what features contributed heavily to a PPP loan being given to an individual living in a LMI community. Since each borrower in our dataset was classified as a 0 (not belonging to a LMI community) or 1 (belonging to a LMI community), we ran one model, a classification random forest model.

B. Classification Random Forest LMI Model:

Once again, we ran 5 different random forest models, each with a different number of trees (100, 200, 300, 400, 500) to split the data. Within each model, we ran a random forest model on "LMI" using our training data, changing the number of trees in each model. Once our model ran, we proceeded to make predictions using our established model on our test data, comparing our predictions to the actual test values, creating an accuracy rate from the results. As shown below, we chose the model with 100 trees as our final model since it had the highest accuracy rate (69.864%).

Figure #8: Random Forest Accuracy By Tree (LMI Prediction)

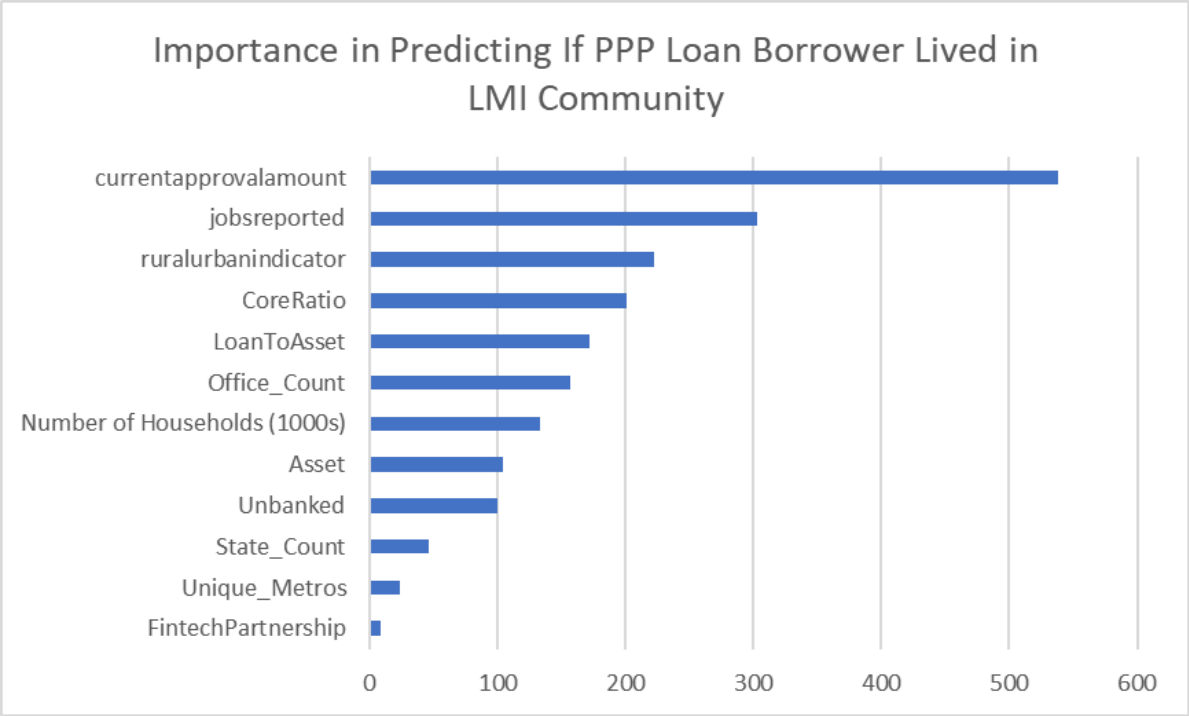
Random Forest Models	
Number of Trees	Accuracy
100	66.73%
200	66.76%
300	66.83%
400	66.77%
500	66.77%

*Derived from File 14

C. LMI Model Results:

To see how our independent variables, including fintech sponsorship impacted "LMI Indicator" we printed the importance report for our final model. As shown within the importance graph, "currentapprovalamount" and "jobsreported" had the greatest impact on whether a PPP borrower lived in a LMI community. These results make sense because the amount of money a business needs to pay for their employees and the amount of people they employ would contribute heavily to the economic prosperity of where they are located. Looking specifically at the fintech partnership variable, we once again found the fintech variable had the least amount of importance when it came to predicting our dependent variable. Therefore, whether or not a community bank partnered with a fintech had relatively low importance to predicting if a borrower lived in a LMI community.

Figure #9: Random Forest Variable Importance (LMI Prediction)



*Derived from File 14

D. LMI Model Importance:

It shows that fintech partnership for community banks isn't a successful way for them to reach lower-income borrowers more so than not partnering with a fintech company. Once again, this could be because a community bank may already have a loyal customer base therefore a fintech partnership may have only reached their existing customers. Instead, 'currentapprovalamount' and 'jobsreported' are the two most important predictors in predicting LMI status of the loan borrower. These two factors go hand-in-hand; if the number of jobs reported is low, loan applicable amounts will be low, and vice versa. Fintech partnership may be independent of loan borrower's LMI status, contrary to our null hypothesis. In order to increase LMI community outreach through fintech partnership in future economic downfalls, community banks will need to analyze their previous tactics used during Covid-19.

VIII. Further Discussion & Recommendations:

A. Further Data Gathering and Cleaning

Given the limitations of data collection and cleaning methods, our key factor of interest can only be identified as a dummy variable. Had more information surrounding the topic been available publicly, it would have provided a more sophisticated understanding of the impact that fintech partnership with a community bank had on PPP loans. For example, if the date of loan application and loan approval were available, it would have allowed us to calculate the average time it took for a community bank and fintech-partnered bank to process the application and measure level of efficiency quantitatively. As each passing day would have been arduous for many small-to-medium businesses during the standstill pandemic period, a quick processing time would have been a great indicator in determining how well the PPP loan policy has been

implemented to reach all corners of the nation. With such helpful metrics to prove an efficient partnership, the government can take advantage of fintech partnership for other policy deliveries, such as distribution of stimulus checks, collaboratively.

B. Web scraping

Due to lack of data surrounding fintech partnerships with community banks, our team resorted to using a web scraping method to list partnerships manually. This was the only way to exhaustively list the partnerships for this analysis. However, it is likely that some partnerships may not have been captured in the scrapings. Some banks or fintechs simply may not have announced their partnership online, which makes it impossible for the scraper to parse through using the keywords that we decided to use, such as “PPP loan”, “Fintech”, and “Community bank”. If the partner’s website or post only hints collaboration in a nuanced manner, such as posting an image or video that announces partnership rather than using html, they might have fallen out of the radar, whose impact in distribution of PPP loans will be unknown and unaccounted for.

C. Minority prediction using ANN classifier

A key part of our initial hypothesis is that the fintech partnership helped to serve the minority with its low entry barrier for usage. To prove such correlation between the variables, our team decided to fill in the null values in the minority column that was provided by CSBS using the CNN model that makes use of Artificial Intelligence techniques. As the model’s accuracy was capped at approximately 72%, some minority predictions may be inaccurate, which is a source of noise into the dataset.

D. Neo-bank classification

With the boundaries between traditional and online banking blurring, a new format of banking has been introduced to the market, also known as a neo-bank. A neo-bank is a bank that offers services similar to those of a traditional bank but only in digital space. Therefore, it is neither a community bank nor a fintech. Since this new type of banking colluded with our hypothesis and FDIC had yet to decide on its classification, we decided to drop neo-bank from our analysis. Once there is an update on neo-bank classification, conducting a hypothesis testing on banking efficiency between a fintech and a neo-bank would be beneficial in providing roadmaps for financial institutions and policies.

IX. Policy Suggestions:

PPP loans received an overwhelming response as this was one of the few ways to keep the businesses afloat in an especially difficult time. Had the applications been directed to one application center or bank across the country, it would have created a bottleneck in the process. However, thanks to the presence of community banks, PPP loan applications were dispersed and could be approved in a timely manner. Although fintech partnership was not a significant factor as illustrated by our analysis, it is worth noting that the partnership may help with acceleration of digitalization of the banks. Digitalization is almost essential after experiencing a pandemic like Covid-19 to allow the economy to run smoothly whilst minimizing human interaction to minimize exposure to various risks. However, some people have flagged that digitalization without establishing security measures and/or compliance protocols may provide a platform for scam, as found by some initial studies for PPP loans. Nonetheless, partnerships with fintech still

provide many advantages that cannot be offsetted by this flaw, which can be fixed with enhanced monitoring systems.

X. Conclusion:

Working with real data posed a lot of difficulties in how to best use the dataset in order to answer the question of interest without following potential biases too closely. Our null hypothesis, community bank's partnership with fintech helped to reach minority and LMI population with PPP loans in 2020, is rejected after a careful analysis. As with most financial datasets, our cleaned dataset was still highly skewed. Fortunately, random forest models do not make any assumptions about the variables and dataset and yet yield the highest accuracy. So we conclude that random forest was the best model to test our hypothesis and explain implications surrounding this dataset at 68% on average for two types of borrower populations. The next possible study is to investigate the impact had by the fintech partnership overall.

XI. Appendix:

- Team Github: **https://github.com/mgd2448/CSBS-Data-Analytics-**
- File 0: CSBS Data Analytics 2022 - Proposal.docx (First Proposal)
- File 1: Bank_Listing_Current_Quarter.xlsx (Community banks: Number by state and asset size. Banking Strategist. (2021). Retrieved April 23, 2022, from https://www.bankingstrategist.com/community-banks-number-by-state-and-asset-size
- File 2: current-community-banking-reference.csv (Dataset provided by CSBS)
- File 3: currentcommunitybankingreferencedata-notestousers.pdf (Notes for File 2 provided by CSBS)
- File 4: ppp-data-dictionary.xlsx (Data dictionary for File 2)
- File 5: Test_Wesbcraper.py (Web scraper model to find fintech partnership in Python)
- File 6: Minority_Prediction_Saved_Model_Commented.py (Trained weights for Python ANN model)
- File 7: Minority_Prediction_Applied_Model_Commented.py (ANN model used to predict minority status in Python)
- File 8: Minority_Prediction_FilledIn.zip (Results in CSV from File 8)
- File 9: Community_Bank_Data_Commented.yxmd (Alteryx workflow)
- File 10: Full_Model_Data.zip (Our cleaned data zip file containing full dataset in CSV)
- File 11: Train_Model_Sample.zip (Our cleaned data zip file containing train dataset in CSV)
- File 12: Test_Model_Sample.zip (Our cleaned data zip file containing full dataset in CSV)
- File 13: Final Analysis.R (Logistic Regression and Random Forest model analyses)