Analysis of the Paycheck protection program from a socio-economic lens

IC#22031 - Abhijit Haridas, Drishti Jain,
Girish Saraf, Pratyush Gupta Uddagiri

The problem

- Removal of applications from the loans dataset.
- Necessary information not provided to lenders?
- Lender cancellation after obtaining necessary information
- Fraudulent applicants?
- Why?

The Process

Clean data

Fixing zipcodes , categorical variable creation

Process data and merge it

With NAICS and Census data

Visualize with Tableau

Demographic feature comparison

Feature Selection & Imbalance reduction

Feature selection using backward regression and SMOTE for class imbalance

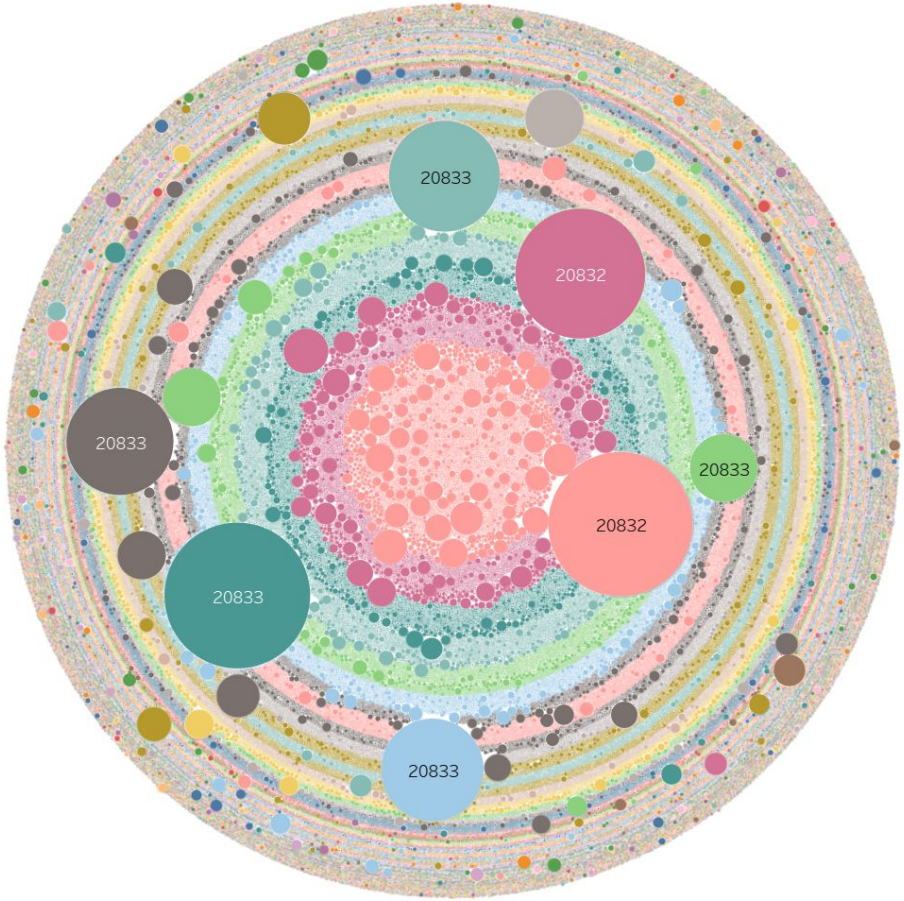
Prediction

Various supervised classification ML models

Model evaluation and tuning

Hyper parameter tuning for generalizability

Loan Applied Amount- Top 30 Lenders



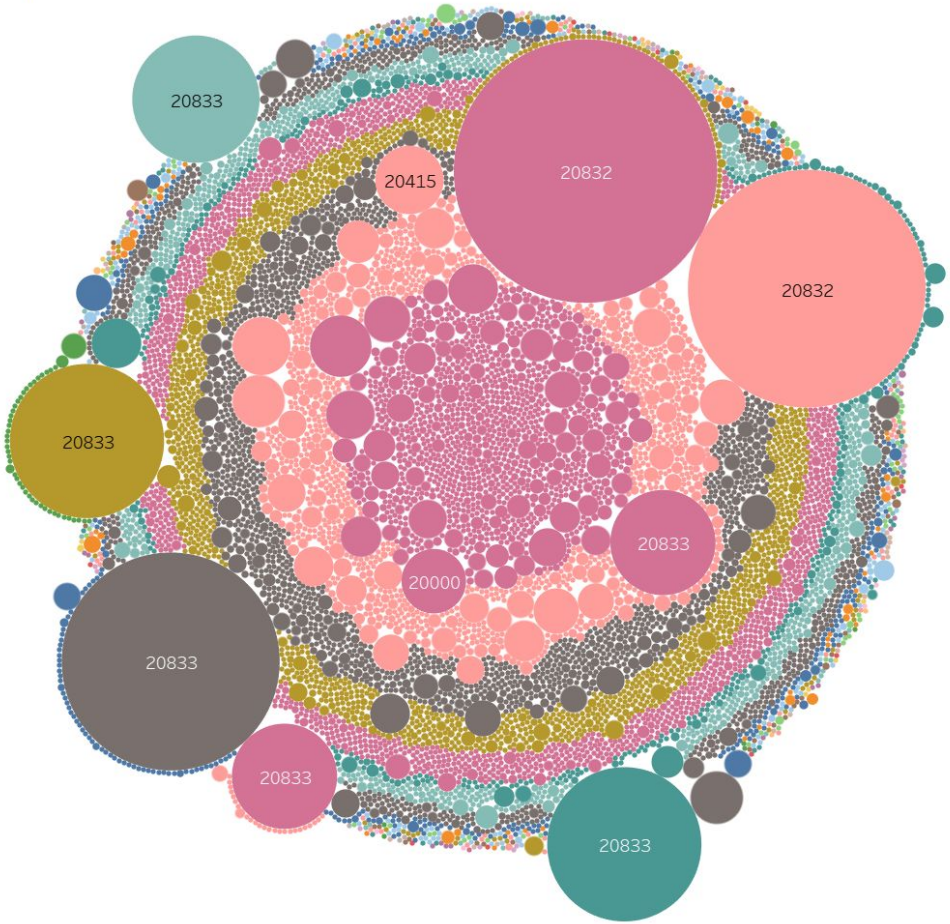
Is Removed

No

Lender

- Prestamos CDFI, LLC
- Capital Plus Financi..
- Harvest Small Busi..
- Benworth Capital
- Cross River Bank
- Fountainhead SBF L..
- BSD Capital, LLC db..
- Bank of America, N..
- Itria Ventures LLC
- Kabbage, Inc.
- Truist Bank d/b/a B..
- Wells Fargo Bank, N..
- Customers Bank
- Synovus Bank
- Celtic Bank Corpora..
- United Community ..
- Ameris Bank
- Regions Bank
- Readycap Lending, ..
- JPMorgan Chase Ba..
- WebBank
- South State Bank, ..
- Renasant Bank
- MBE Capital Partne..
- Amur Equipment Fi..
- Square Capital, LLC
- Queensborough Na..
- Colony Bank
- Pinnacle Bank
- PNC Bank, National ..
- Bank OZK
- United Bank
- Leader Bank, Natio..
- First Horizon Bank
- First-Citizens Bank ..
- The Enterprise Cent..

Loan Applied Amount- Top 30 Lenders



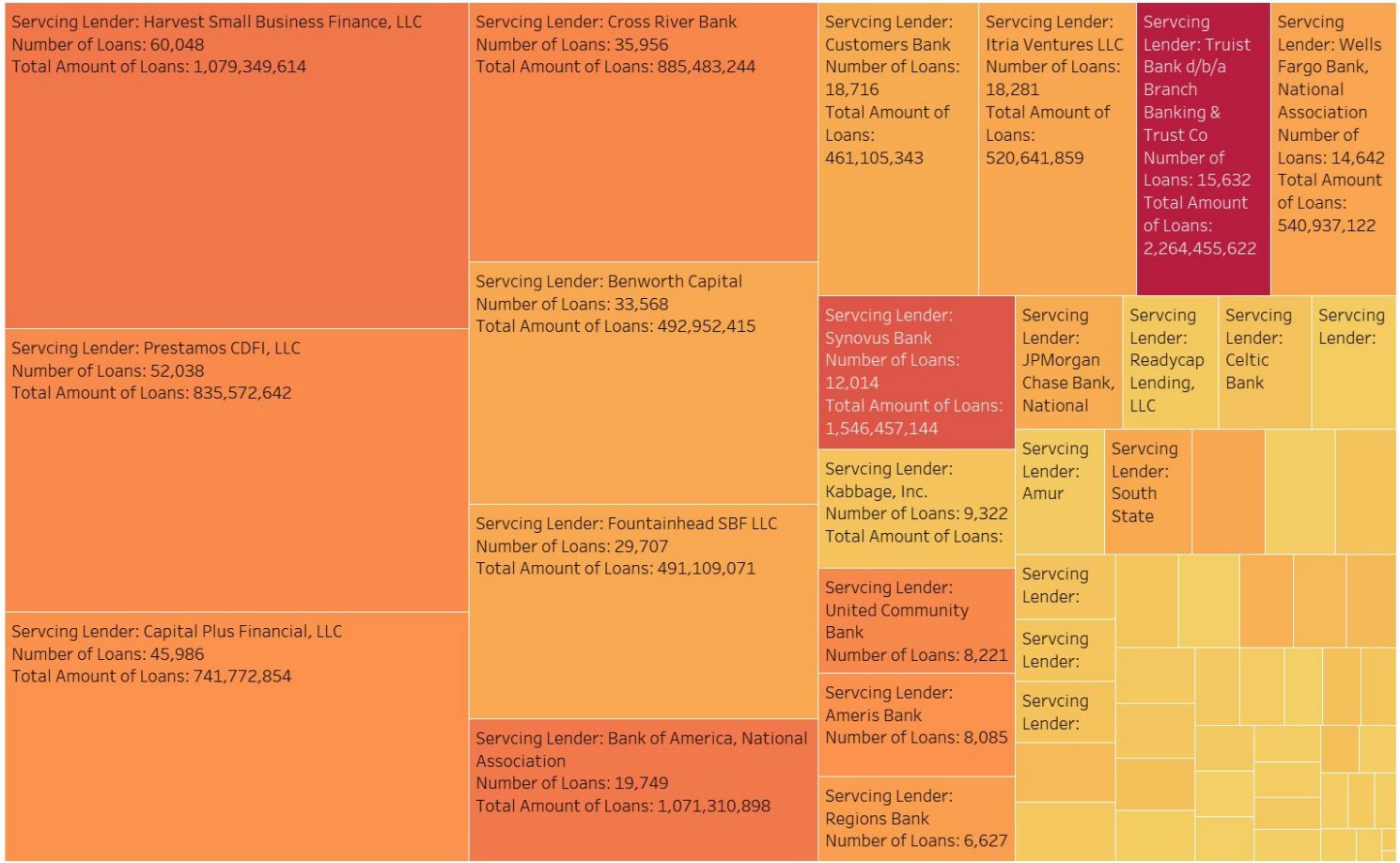
Is Removed

Yes

Lender

- Capital Plus Financi..
- Prestamos CDFI, LLC
- BSD Capital, LLC db..
- Customers Bank
- Leader Bank, Natio..
- Harvest Small Busi..
- Benworth Capital
- Itria Ventures LLC
- Readycap Lending, ..
- Texas National Bank
- Newtek Small Busin..
- The Enterprise Cent..
- 1st Choice CU
- American Lending C..
- Blue Ridge Bank, N..
- Cross River Bank
- Amur Equipment Fi..
- Access to Capital fo..
- MBE Capital Partne..
- Fountainhead SBF L..
- Heritage Bank, A Di..
- FC Marketplace, LL..
- Synovus Bank
- DreamSpring
- South State Bank, ..
- BankSouth
- Northwest FCU
- WebBank
- MVB Bank, Inc.
- Navy FCU
- Citizens Trust Bank
- Community Federal ..
- Hope Enterprise Co..
- Peach State FCU
- Kabbage, Inc.
- Ameris Bank

Servicing Lenders- Number of Loans and Amount



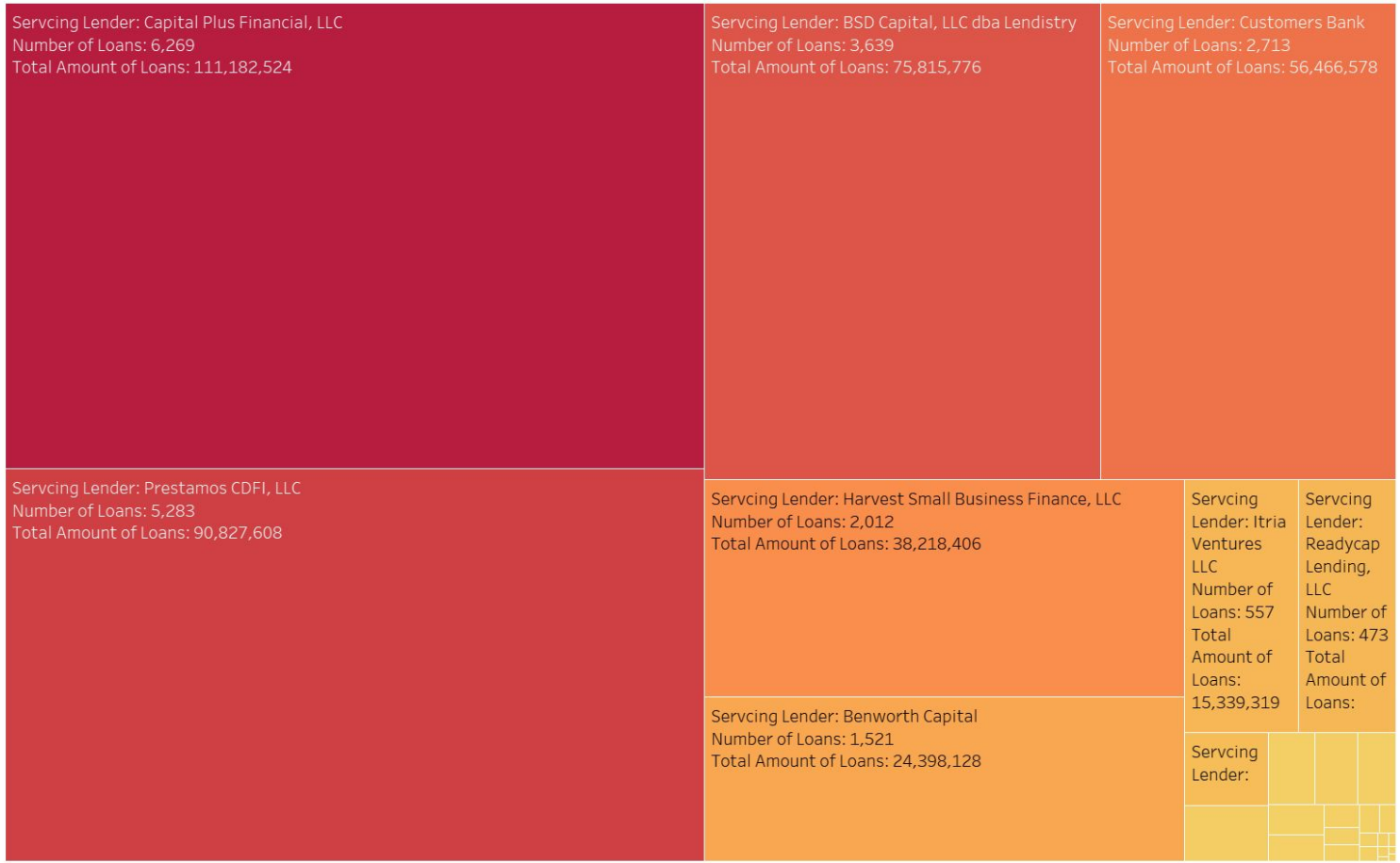
Is Removed

No

SUM(Amount)

32,565,840 2B

Servicing Lenders- Number of Loans and Amount



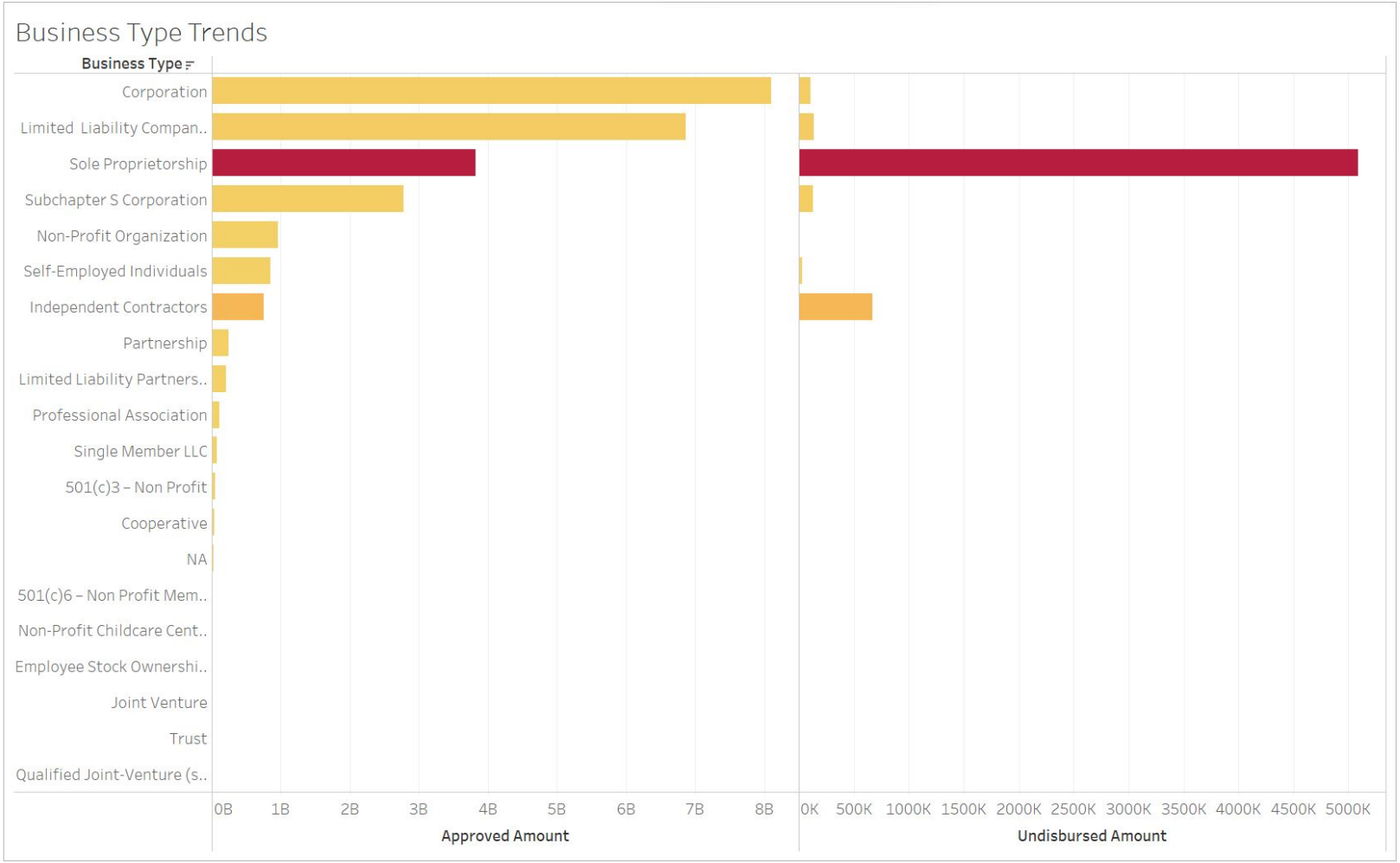
Is Removed

Yes

SUM(Amount)

2,000 111,182,524

Business Type Trends



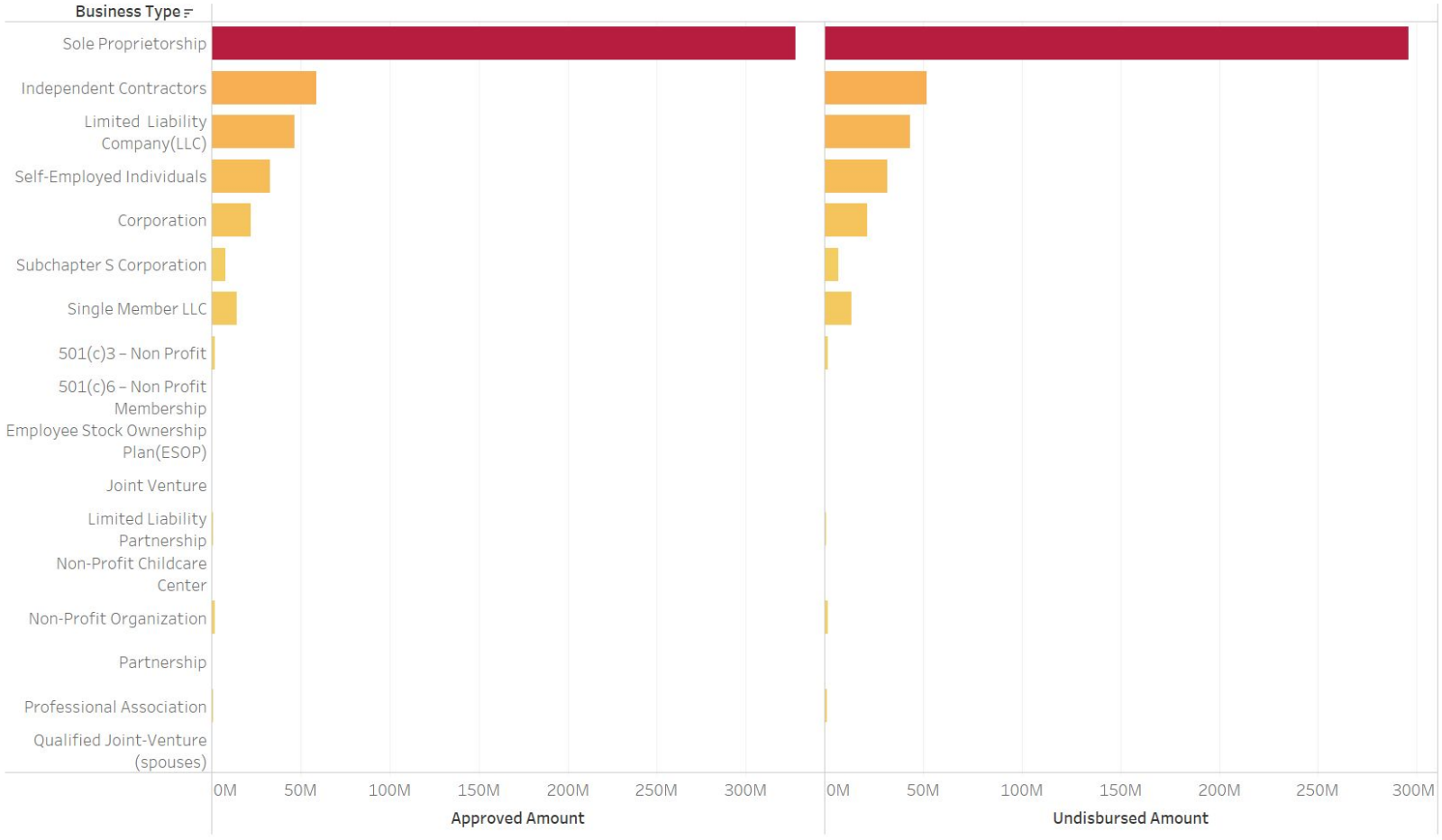
Is Removed

No

SUM(Undisbursed Amo...

0 5,084,992

Business Type Trends



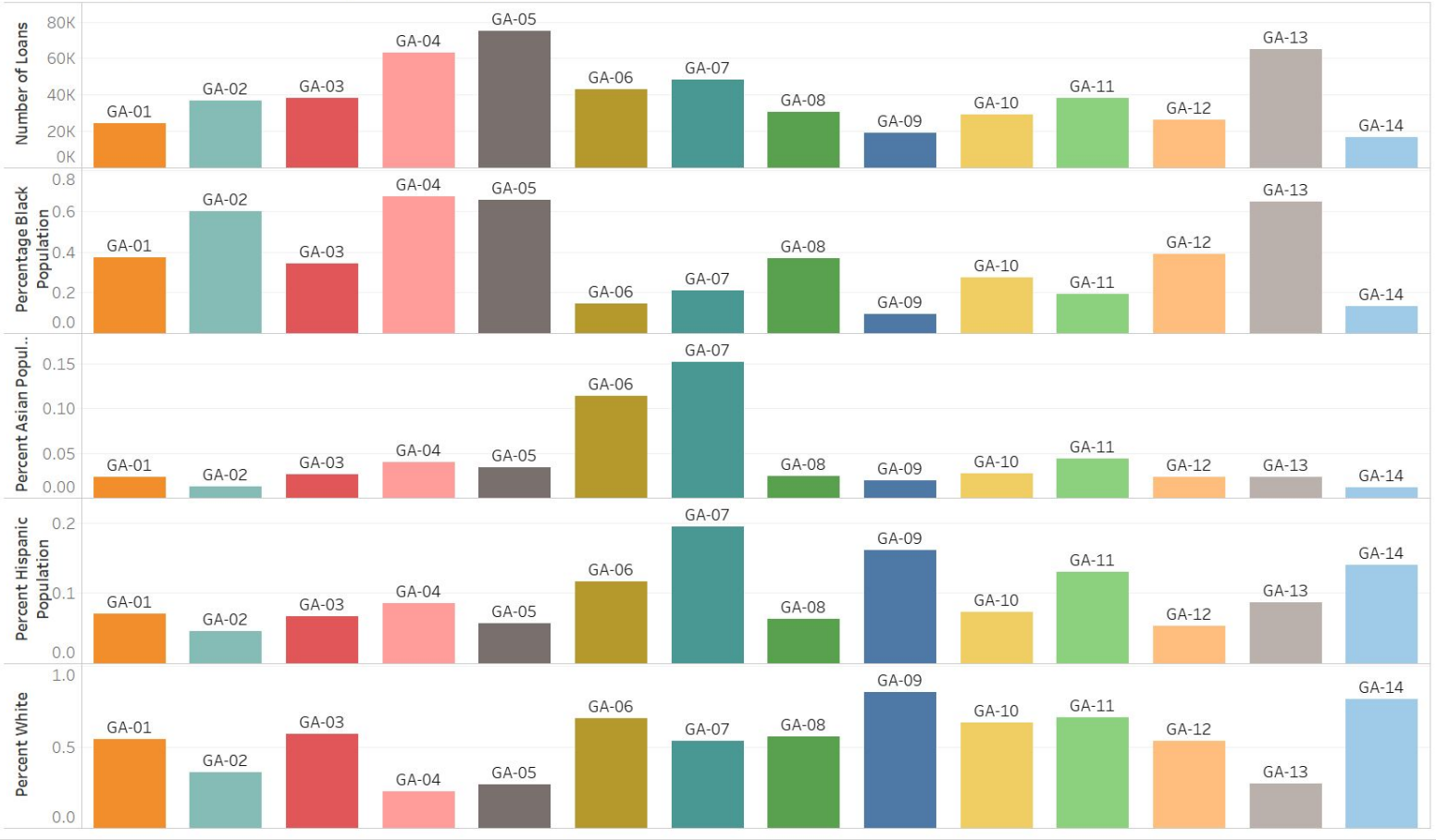
Is Removed

Yes

SUM(Undisbursed Amo...

2,103 295,702,791

Loans Approved Based on Zip Demography



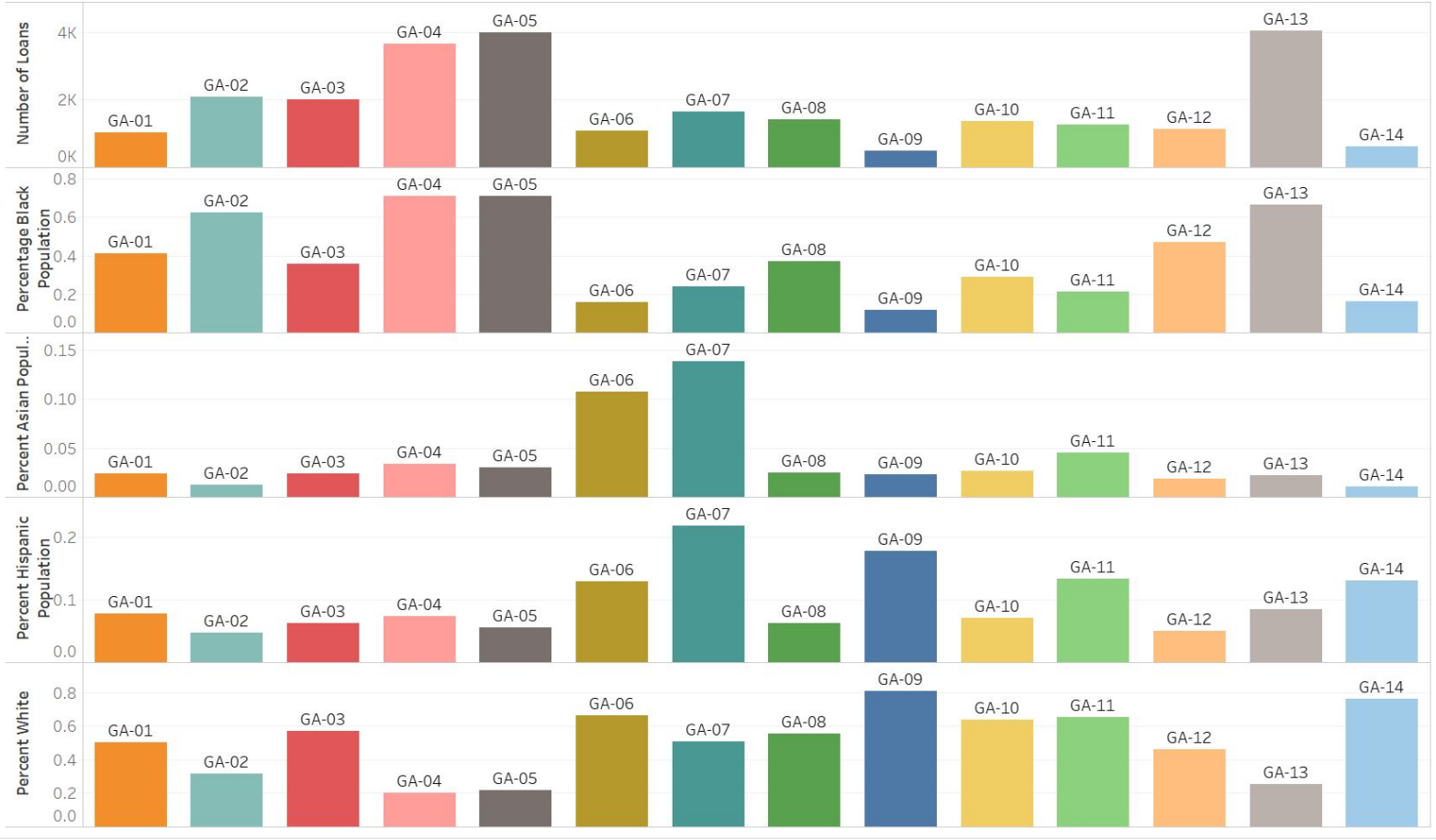
Is Removed

No

Congressional District

- GA-01
- GA-02
- GA-03
- GA-04
- GA-05
- GA-06
- GA-07
- GA-08
- GA-09
- GA-10
- GA-11
- GA-12
- GA-13
- GA-14

Loans Approved Based on Zip Demography



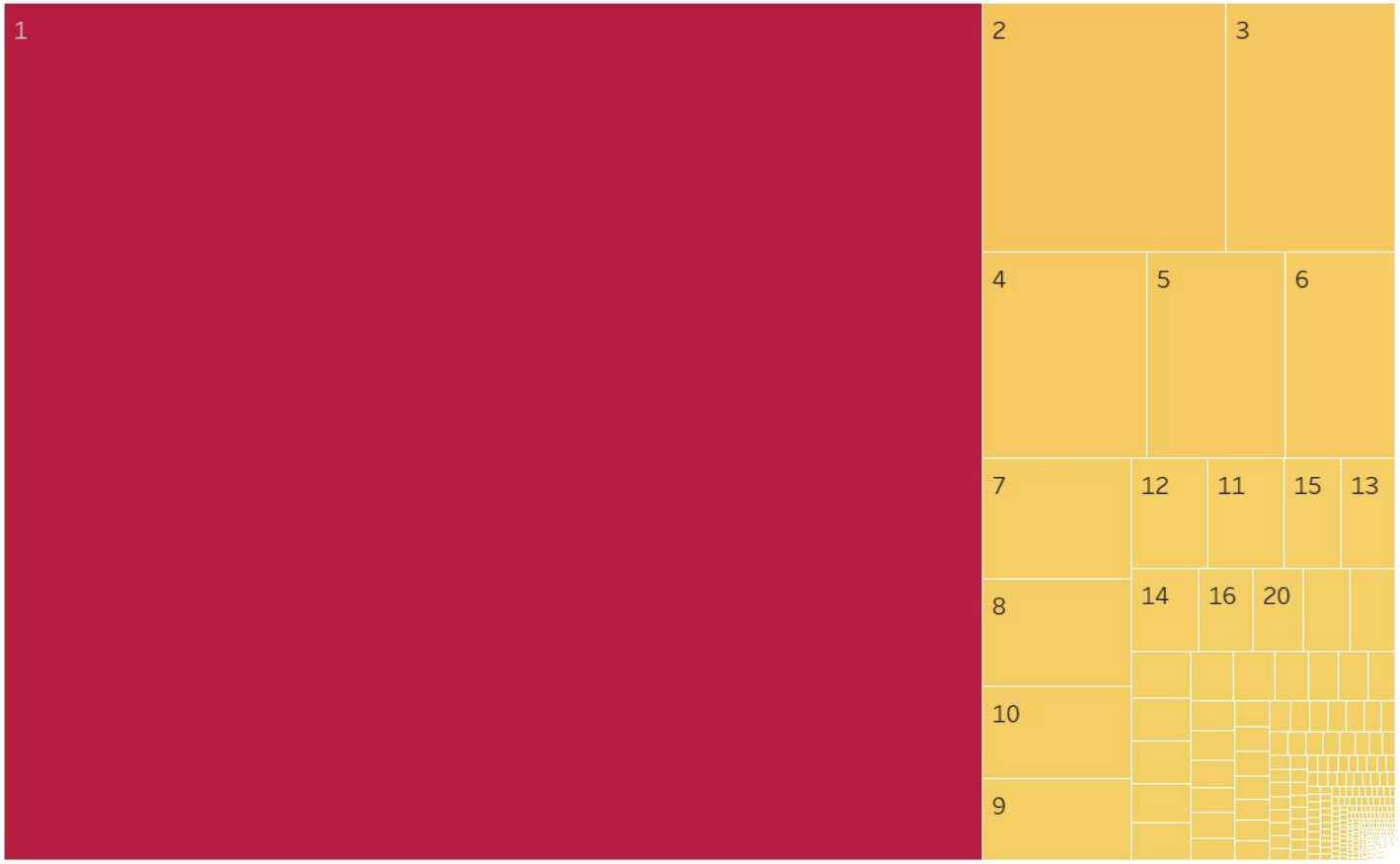
Is Removed

Yes

Congressional District

- GA-01
- GA-02
- GA-03
- GA-04
- GA-05
- GA-06
- GA-07
- GA-08
- GA-09
- GA-10
- GA-11
- GA-12
- GA-13
- GA-14

Jobs Retained Trends



Is Removed

No

CNT(Loan Number)

1 3,89,415

Jobs Retained Trends

Is Removed

Yes

CNT(Loan Number)

1 24,836

Machine learning models -

- To understand variable interactions better, we used Tableau to visualize the data and gain insights.
- High class imbalance was observed in the case of loan amount and ‘jobs retained’, SMOTE was used to systematically introduce synthetic data points and reduce the effect of the imbalance.
- Using these insights and by ranking feature importance, we built our models using various classification techniques such as Logistic Regression, Random Forest Classification, Decision Tree classification, etc.
- Evaluation was done based on precision, recall, confusion matrix and AUC.
- Finally, we tuned the selected model using hyper-parameter tuning to get the best possible generalizable results.

Results -

	Modelling Algorithm	Accuracy	Precision	Recall	F1 score	ROC AUC score
1	LogisticRegression	87.57%	20.91%	64.10%	31.54%	76.39%
2	Random Forest Classifier	87.52%	21.49%	67.68%	32.63%	78.06%
3	Decision Tree	83.63%	19.18%	82.98%	31.16%	83.32%
4	Gradient Boosting Classifier	86.97%	21.90%	74.71%	33.87%	81.13%
5	Gaussian NB	71.53%	13.53%	99.77%	23.84%	84.99%

Post Hyper Parameter Tuning of Decision Tree Classifier:

Modelling Algorithm	Accuracy	Precision	Recall	F1 score	ROC AUC score
Decision Tree	86.31%	21.21%	76.00%	33.15%	81.39%

Conclusion

The aim of this project was to understand the features that were specific to loans that were removed and how they differed from the the existing loans in the PPP database. A few insights from exploratory data analysis are as follows:

- Majority loans(About 95% of total count) being approved for the forgiveness cap amount-$20,833 with a Truist Bank having the highest dollar amount of loans provided.

- Majority of the removed loans applications(About 95% of total count) were retaining **1** job.

- Sole proprietorship applicants had the highest number of rejected loans in comparison to Corporations being the highest number of loan applicants.

- Loans applied through GA-13 congressional district composed of the majority of the removed loans dataset. GA-13 also had the a majority percentage of black population when compared to overall population.

On training the entire dataset with various binary classification models, a decision tree classifier best converged and explained the validity of prediction of whether a loan would be part of the removed loans dataset.
False negatives were to be minimized since there was a high imbalance between the size of removed loans and existing loans. Hence “recall” was the evaluation metric to be optimized.

Thank You!

Q&A?