Georgia Tech
Scheller College
of Business

Acorns: Paycheck Protection Program Analysis

April 19, 2023

Meet the Team

Paula Simonetti

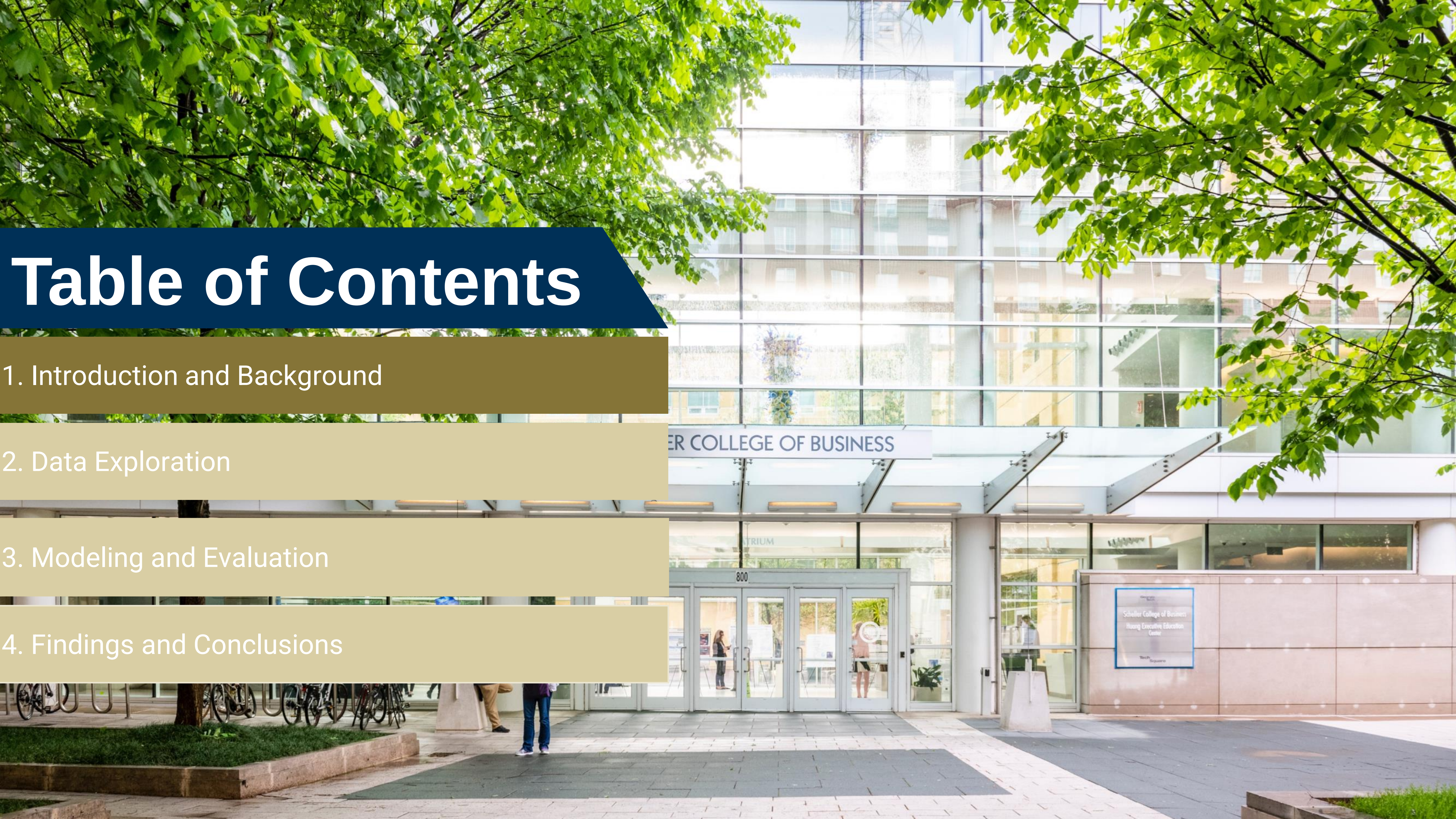
Emma Watkins

Zach Labkovski

Cassidy Gasteiger

Omar Mikawi

Table of Contents



Our Goal

Understand the need for Small Business Association Paycheck Protection Program loans based on labor market data so loan providers can better prepare for a future economic shutdown event

IMPACT ANALYSIS

Assess impact of the pandemic shutdown by analyzing county and industry impacts in Georgia

IMPACT FORECAST

Predict impact of the pandemic shutdown in terms of job loss based on pre-shock county attributes

DEMAND FORECAST

Predict number of PPP loans received by county and total dollar amount of loans received by county based on county attributes

Business Case

PPP loans:

By understanding the impact of jobs lost and the demand for loans, we can forecast future need, if another pandemic were to occur.

Why Acorns cares:

Acorns does not currently offer loans as part of their services. However, start-ups such as Blueacorn and Womply provided support for small businesses needing PPP loans and reaped more than $3 billion from fees.(1) Acorns targets a similar underserved market and if properly prepared could reap the same benefits in the case of a future government loan program.

1) https://www.nytimes.com/2021/06/27/business/ppp-relief-loans-blueacorn-womply.html

Table of Contents

Literature Review: Labor Impacts

Key Impact Factors

—

- Worker classification:
  - work-from-home ability
  - essential vs. non-essential
- Wage cuts
- Policy response (i.e., CARES Act)
- Demographic data
- Educational attainment
- Industry (SAIC)
- Recovery rates

Labor Impact Estimation Approaches

—

- Regression models
- Industry ratios based on employment type classification

Key Findings

—

- Georgia lost 609K non-agriculture jobs from Feb 2020–Apr 2020
- Hospitality and Public sector recovery lags
- Higher rates of unincorporated self-employment post-COVID
- Recovery rates differ by segment:
  - Black Georgian's underemployment rose 300 basis points compared to pre-COVID levels
  - White women's un- and under-employment rate also did not recover to pre-COVID levels
  - White and Hispanic men's employment recovered to pre-COVID levels in the same timespan

Labor Impact At-a-Glance

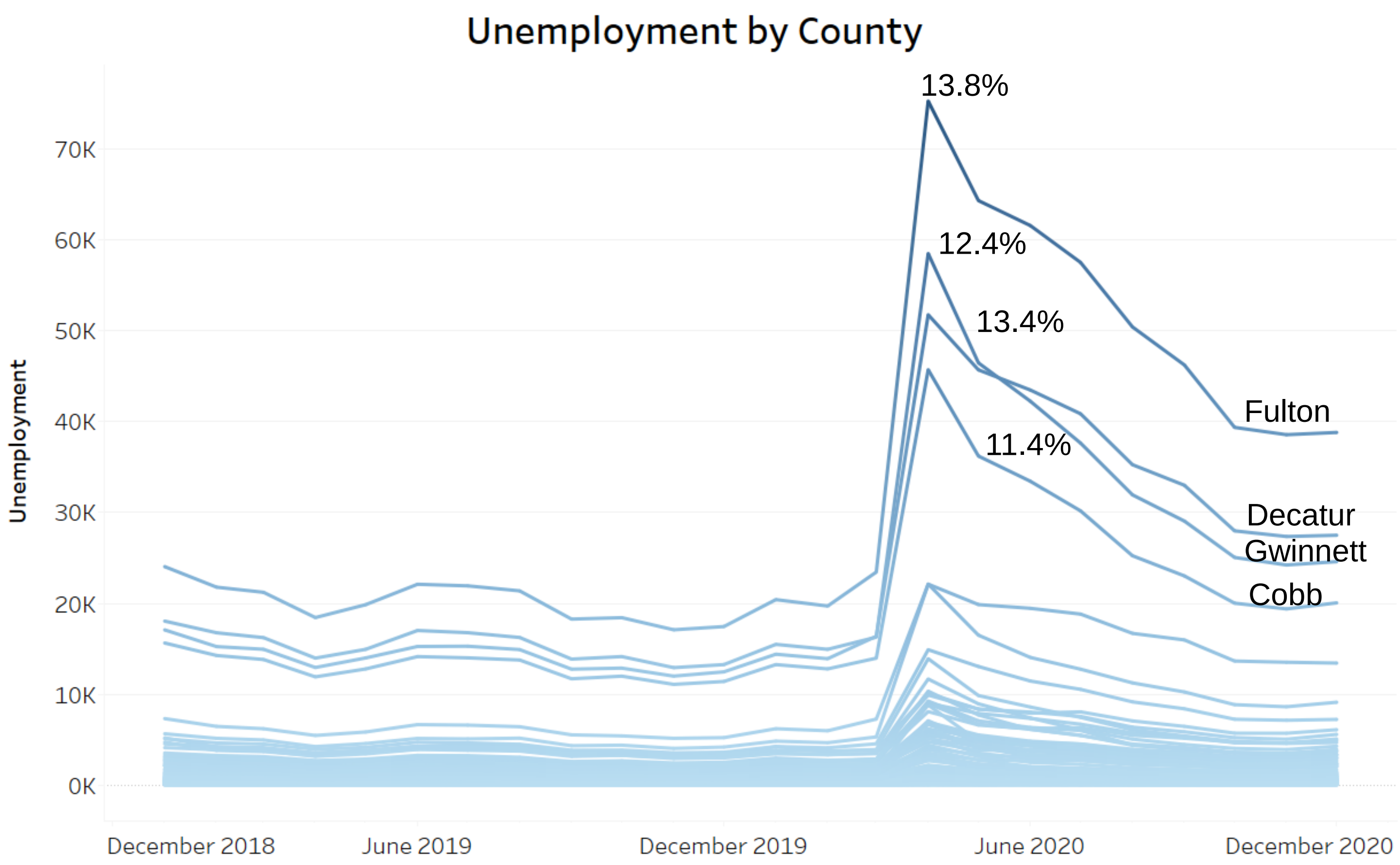
Unemployment spikes in March 2020 across all counties and begins to recover in April 2020

546,995
jobs lost in April

9.2%
jobs lost on average in April

27.1%
of jobs lost in Berrien County (most-impacted)

10.51%
Avg. April 2020 unemployment rate

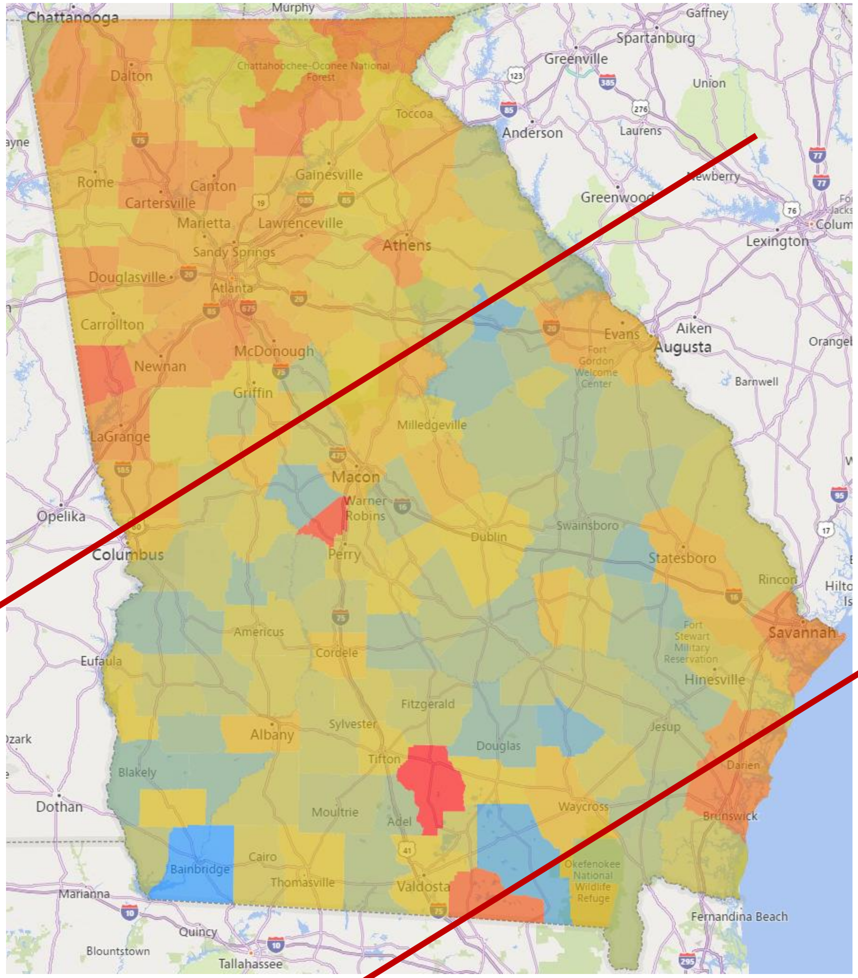


Labor Impact by County

- The space in between the red lines is known as the **"Piedmont Region"** and is represented by smaller businesses and population sizes

- The Piedmont Region experienced **less intense labor impacts** from COVID-19 because of its quantity of agricultural jobs, political leanings, lack of restaurant and retail jobs, and small-town effect.

Heatmap of Job Loss Ratio from March – April 2020 by County



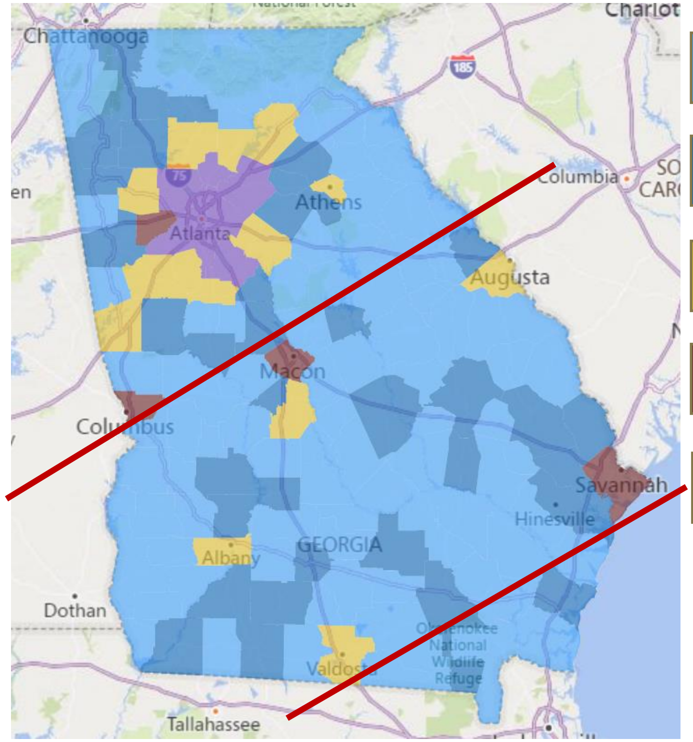
Greatest job loss as percent of total jobs (Berrien County)

Lowest job loss as percent of total jobs (Decatur County)

PPP Loans At-a-Glance

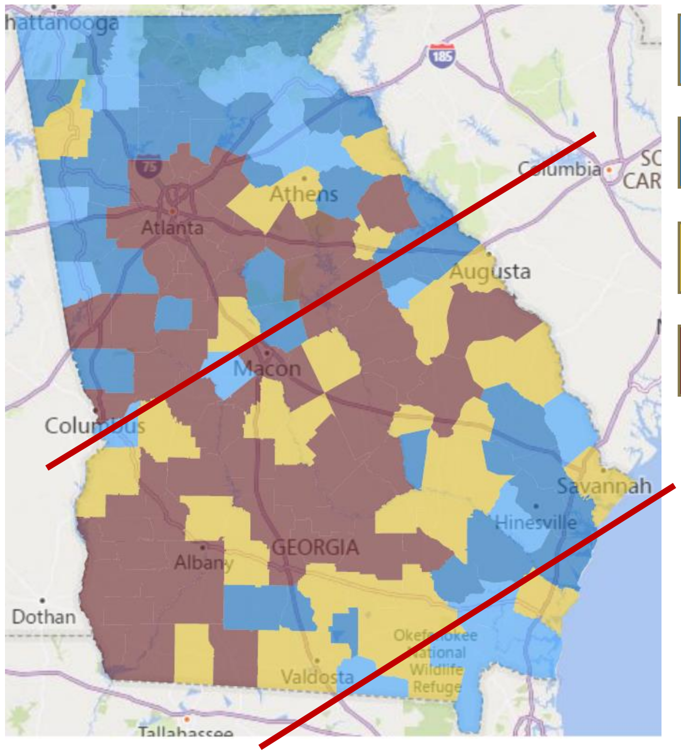
Loans were distributed at the same rate when you compare on a portion of the population basis, but they were smaller loans because a lot of people became self-employed and leaned into their side business.

Count of PPP Loans Issued by County



- < 1,000 loans
- 1,000 – 5,000 loans
- 5,000 – 10,000 loans
- 10,000 – 15,000 loans
- >15,000 loans

Count of PPP Loans Per 10K Labor Force Issued by County



- < 500 loans
- 500 – 750 loans
- 750 – 1,000 loans
- >1,000 loans

$25,724,311,477 in PPP loans granted

584,485 total PPP loans issued

$43,661 average loan amount

4.74 average company size

Labor Impact by Industry

- **Service-providing lost 171k jobs,** more than any other industry
- **Art, Entertainment, and Recreation** has the largest percent change of **-22.4%**
- **Public Administration** and **Utilities** had a **net increase in jobs**
- **Financial Activities** had the smallest percentage change at **–0.7%**

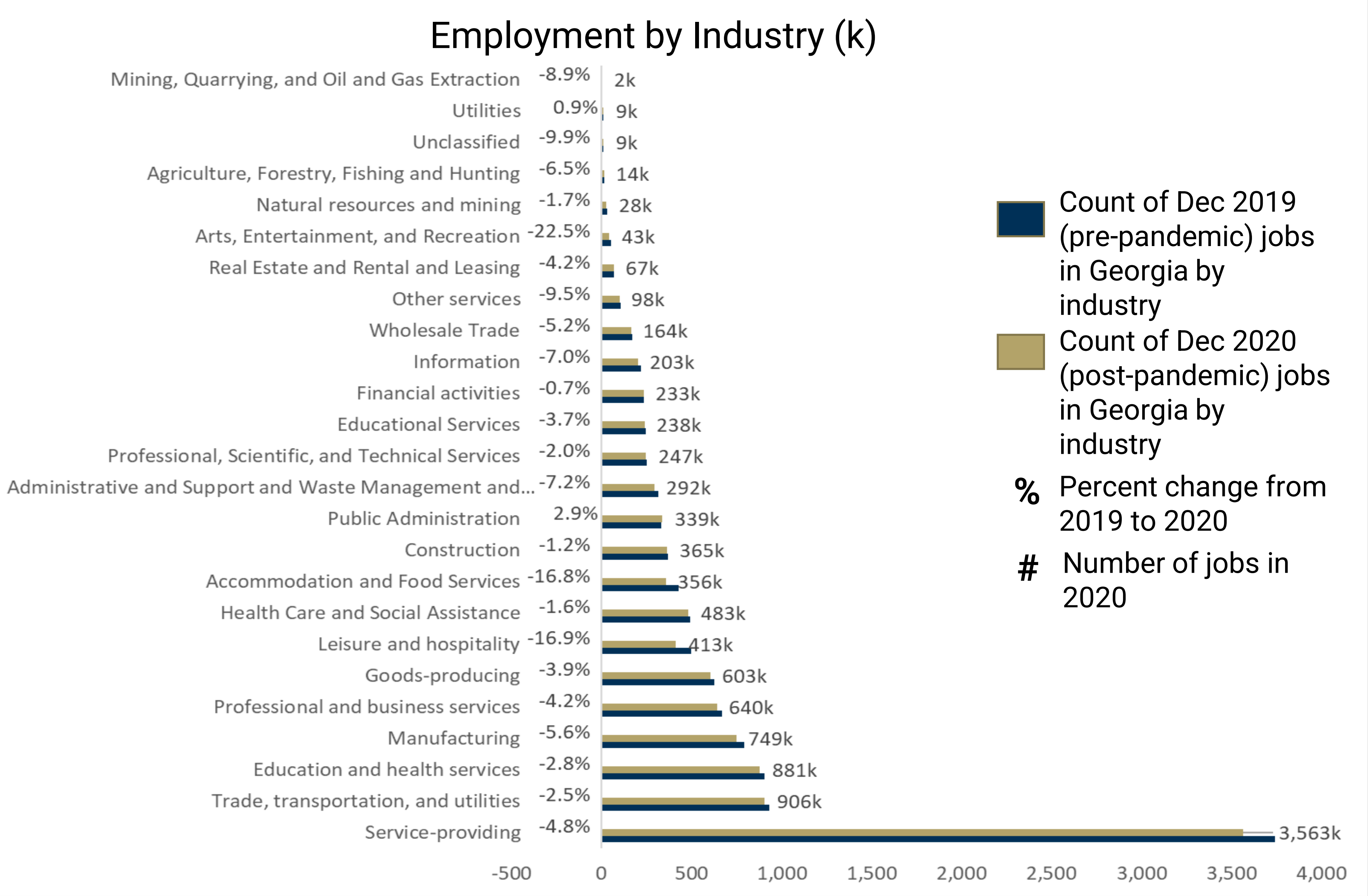


Table of Contents

Modeling Approach

Labor & Loan Impacts

FEATURE ENGINEERING

Separate time series and baseline data and develop secondary variables

- Used March 2020 as baseline (or December 2019 for annualized features)
- Derived ratios for most features and normalized the rest to adjust for population
- Identified y variables to model

FEATURE SELECTION

Remove highly collinear variables and select most-correlated features for further modeling

- Compared VIF scores
- Compared p-values while dropping and adding variables
- Forward feature selection in random forest model
- Lasso regression

MODELING

Test linear regression approaches to build an accurate and interpretable model

- Multivariate linear regression
- Lasso regression
- Random forest regression

Model Features

Explanatory Variables Tested

Population
- Total population
- High population binary flag (if county population > 500,000)

Employment
- Labor force
- Percent of population in the labor force
- March employment (number of people employed in the county in March)
- Unemployment rate (as of year-end Dec. 2019)
- Commuter ratio
- Average working adults/household

Self-Employment
- Self-employment as percent of workforce
- Non-employer establishments as percent of total establishments

Industry Mix
- Total establishments
- Percent of total establishments belonging to each of 31 NAICS industry codes

Wealth Proxy
- Average weekly wages

Target Variables Modeled

Job Loss
- Total number of jobs lost in a county

Loan Demand
- Total number of loans received by a county

Loan Amount
- Total dollar amount received in loans by a county

Non-Employer Establishments

39.9%

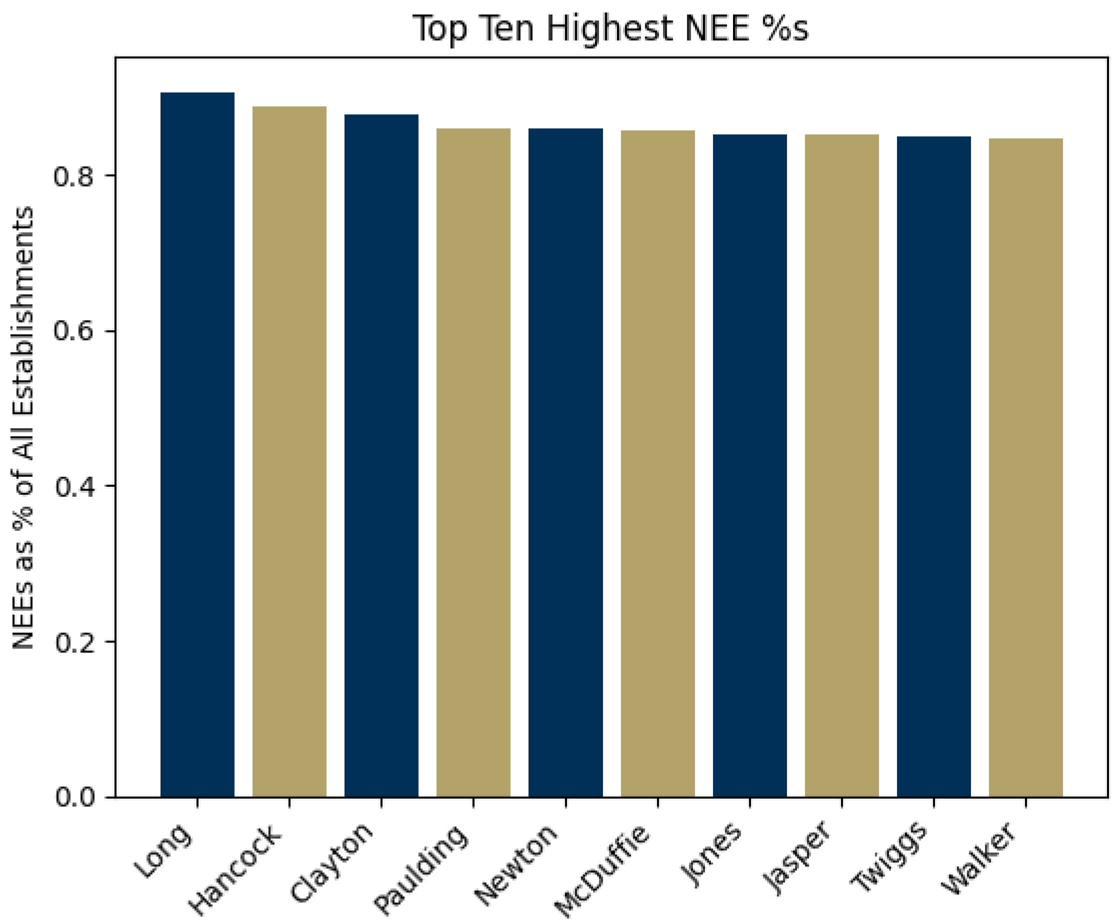
of loans went to NEEs

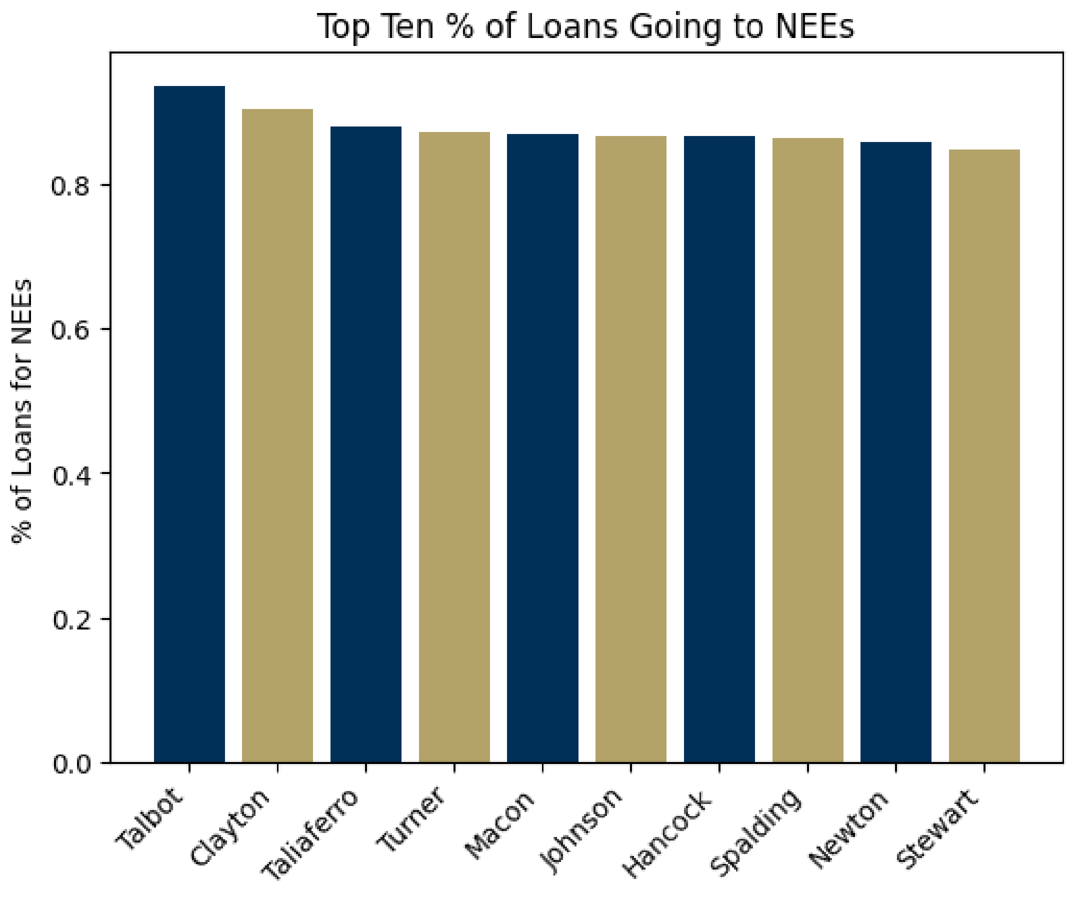
86.2%

issued in Phase II

A non-employer establishment (NEE) is a company that has no paid employees.

This includes gig economy workers and self-employed people.

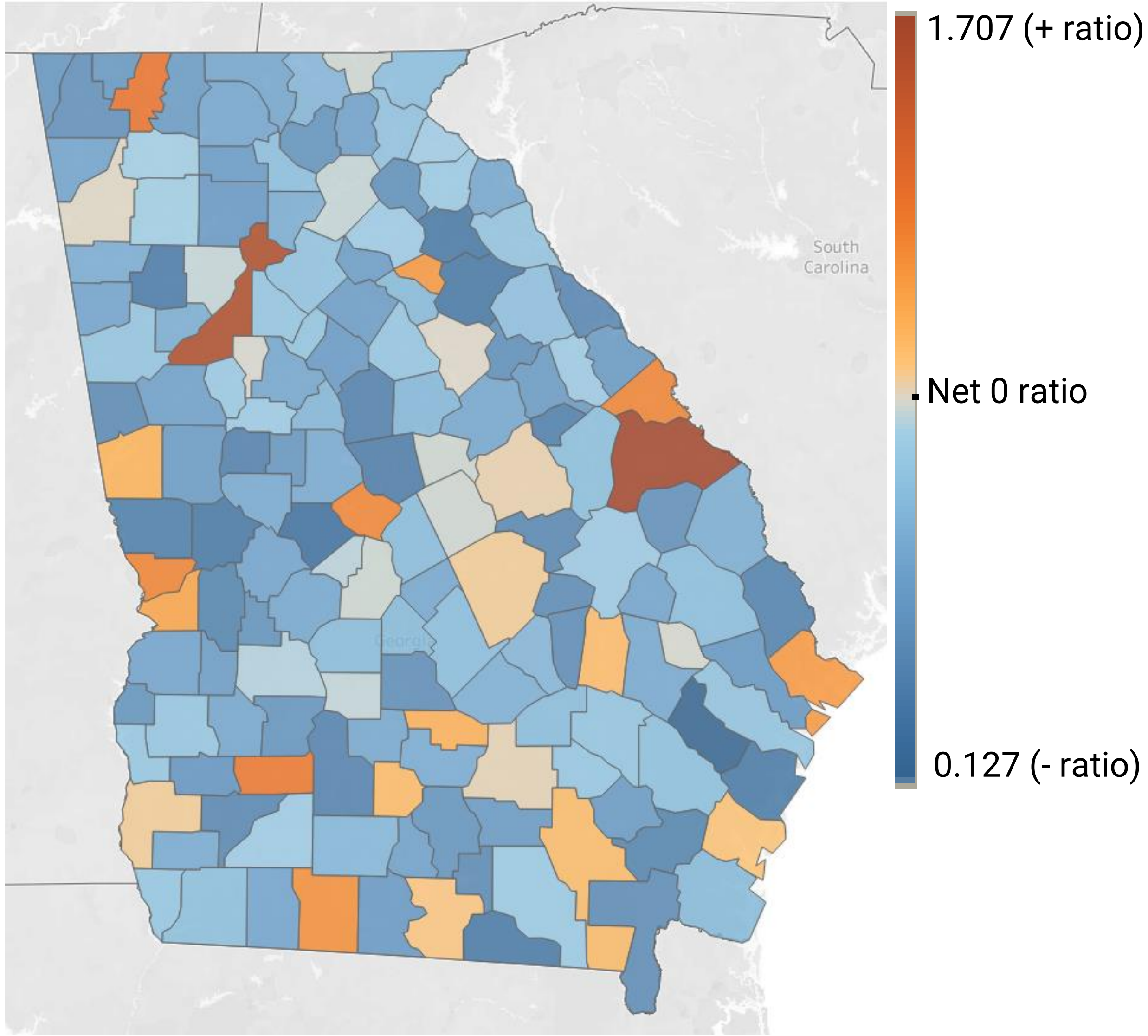




Commuter Ratio

- Local area employment measures number of people employed in a county; employment measures number of people living in a county who are employed
- Taking local area employment / employment gives us a commuter ratio: a sense of how many people are commuting into the county to work
- Counties with a high concentration of jobs (Fulton – Atlanta Metro Area, Chatham – Port of Savannah, Burke – Nuclear Power Plant) have high commuter ratios compared to some rural counties with very low ratios (Long, Crawford)
- Much of the Piedmont Region has a commuter ratio close to 1; most people live where they work

Heatmap of Commuter Ratio by County



Predicting Job Loss

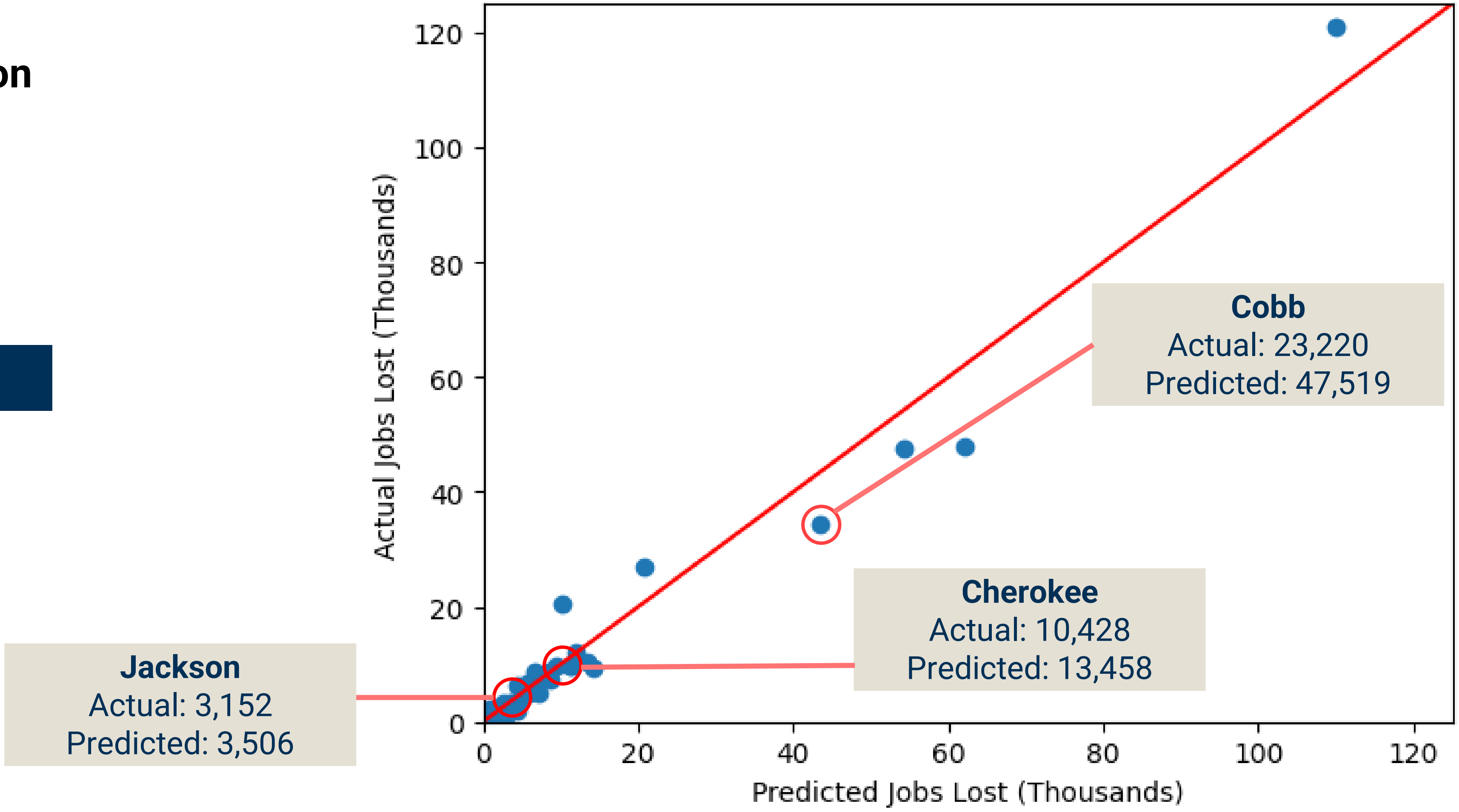
PERFORMANCE

Linear Regression
R2 Score: .963

KEY FEATURES

1. Total Establishments
2. Commuter Ratio

PREDICTIONS



Predicting Loan Demand

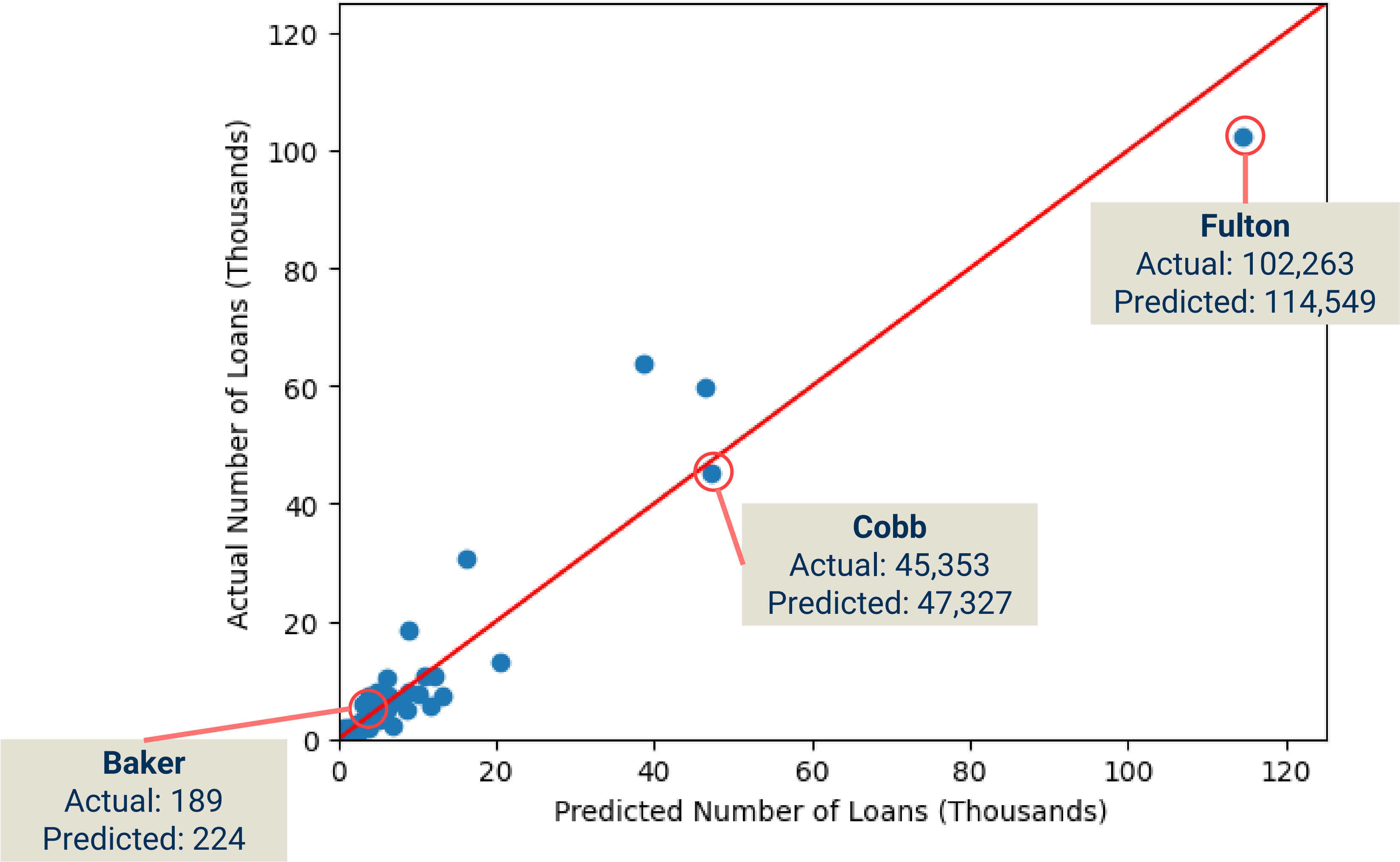
PERFORMANCE

R2 score: .962

KEY FEATURES

1. March employment
2. Self-employment percentage (applied Box-Cox transformation to normalize residuals)

PREDICTIONS



Predicting Loan Amount: Feature Selection

PERFORMANCE

Linear Regression
R2 Score: 0.995

FINAL FEATURES

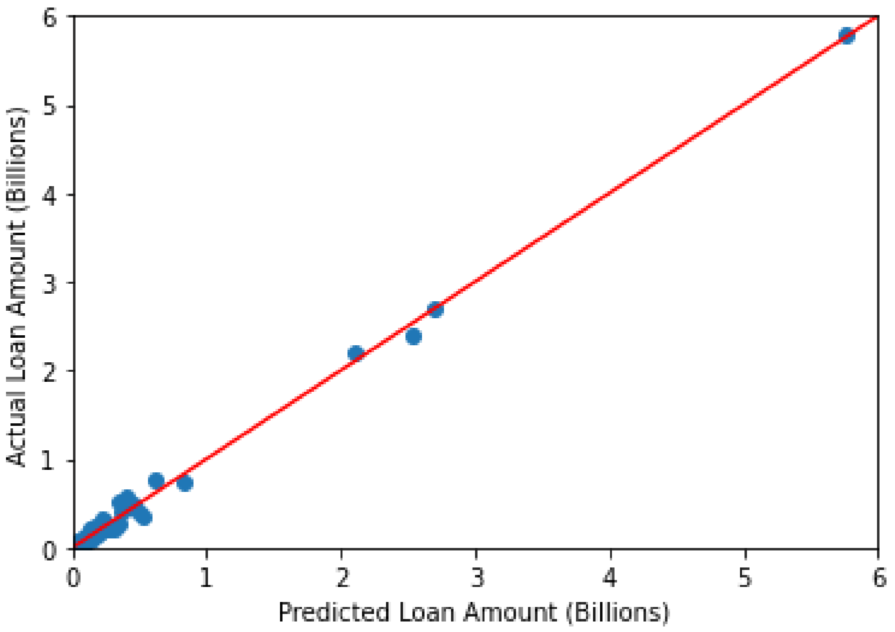
1. March employment (population metric)
2. Non-employer entity rate: NEE/workforce
3. "High Density" County Binary Variable: 1 if pop > 500k, 0 otherwise

Model Features	Adj R2 Score
Employment	0.9877
Emp, NEE/Workforce	0.9897
Emp, NEE/W, Avg Weekly Wages	0.9896
Emp, NEE/W, High Density	0.9944
Emp, NEE/W, HD, HD Interaction	0.9956

Predicting Loan Amount: Model Validation

LINEAR REGRESSION

MAPE: 81.81%



RANDOM FOREST

MAPE: 13.21%

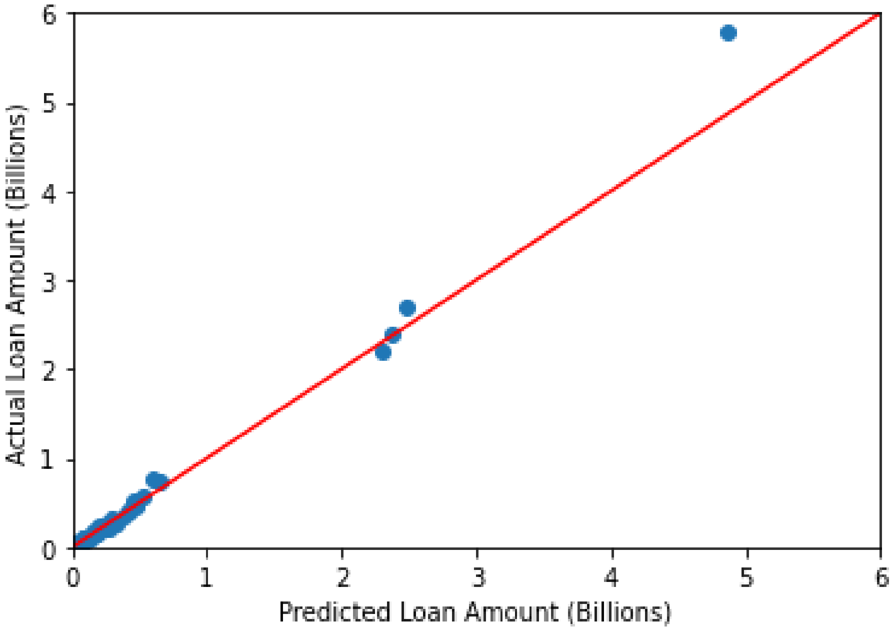


Table of Contents

Key Findings

Job Loss

- A linear regression model can accurately predict job loss in an economic shutdown based on pre-pandemic labor conditions.
- While other variables were highly correlated with job loss by county, the shutdown was so massive and so quick that the only significant predictive variables were population and commuter ratio (that adjusts for the employed in other county, unemployed in-home county disparity).

Loan Demand

- A linear regression model can accurately predict demand for PPP loans by county.
- Demand for loans is mostly a function of population and number of self-employed people, as these are the people who are most likely to need access to loans to keep their small businesses afloat during a crisis.

Loan Amount

- A random forest model can most accurately predict the demand in dollar amount of PPP loans.
- The biggest hurdle to overcome was the difference between high population and low population counties; interaction terms solved this issue.

Business Recommendations

1 COUNTY FOCUS

- Counties with **lower population, high commuter ratio,** and **high NEE percentages** had the highest demand for PPP loans per person
- Targeting these more rural counties in the case of a future government small business loan program could allow Acorns to maximize their reach

2 EARLY ADOPTERS

- **NEEs initially had trouble accessing loans**, as evidenced by the much higher percentage of second-draw PPPs going to NEEs
- Acorns' user-friendly approach makes it an ideal candidate for working with small businesses; it should target NEEs and share information about loan opportunities early to capture this market share immediately

3 GENERALIZABILITY

- Our analysis focused on Georgia, but we expect many of these trends to hold nation-wide
- Further research could test generalizability of our findings to other states

Thank You

Any Questions?

References

Labor Impact References

—

- Osborne Jackson. "An Approach to Predicting Regional Labor Market Effects of Economic Shocks: The COVID-19 Pandemic in New England." *Federal Reserve Bank of Boston.* June 29, 2020.
- Charles Gascon. "COVID-19: Which Workers Face the Highest Unemployment Risk?" *Federal Reserve Bank of St. Louis.* March 24, 2020.
- Ray Khalfani. "State of Working Georgia: Pandemic Job Numbers Show Ongoing Progress at the Surface, but Inequities Persist Below." *GBPI.* May 12, 2022.

Loan Classification References

—

- John M. Griffin, Samuel Kruger, and Prateek Mahajan. "Did Fintech Lenders Facilitate PPP Fraud?" *Journal of Finance.* 2021.
- "Small Business Administration Paycheck Protection Program Phase III Fraud Controls." *Pandemic Response Accountability Committee.* January 1, 2022.
- Office of Inspector General. "Inspection of SBA's Implementation of the Paycheck Protection Program." *U.S. Small Business Administration.* January 14, 2021.

Additional Background

What is a PPP loan?

- Paycheck Protection Program (PPP) is a business loan program established by the United States federal government in 2020 through the Coronavirus Aid, Relief, and Economic Security Act (CARES Act)
- Provides low-interest private loans to pay for small business’s payroll and certain other costs
- The amount of the PPP loan was based on the applicant's payroll costs
- Approximately $800 billion in low-interest uncollateralized loans from April 3, 2020, through May 31, 2021.

Suspect Loan Analysis: Lit Review

Key Findings

- 1.8M of PPP's 11.8M loans showed signs of fraud nationwide
- ~$78B in PPP money taken illicitly
- Most common fraud: business didn't exist before Feb. 2020
- PPP program had very limited review process and relied on lender for all determinations
- Most cases of flagged loans have gone uninvestigated

Key Detection Factors

- Non-registered businesses
- Multiple businesses at residential addresses
- Abnormal compensation per employee
- Inconsistencies with other government programs

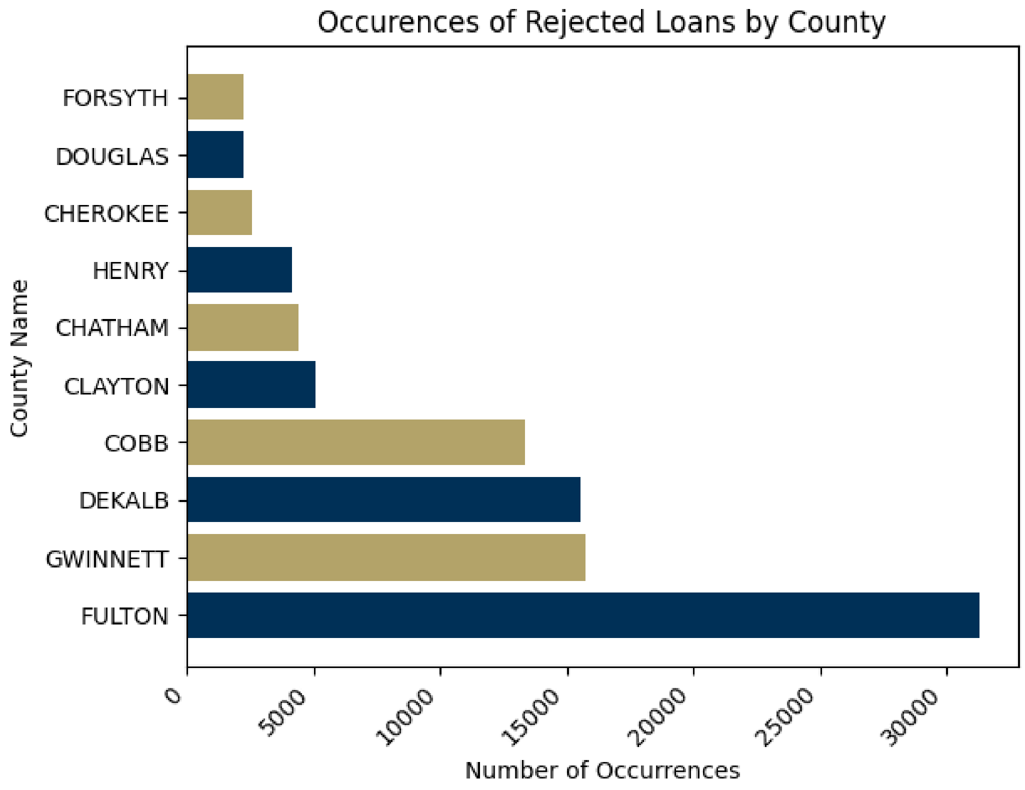
Fraudulent Loan Detection Outcome

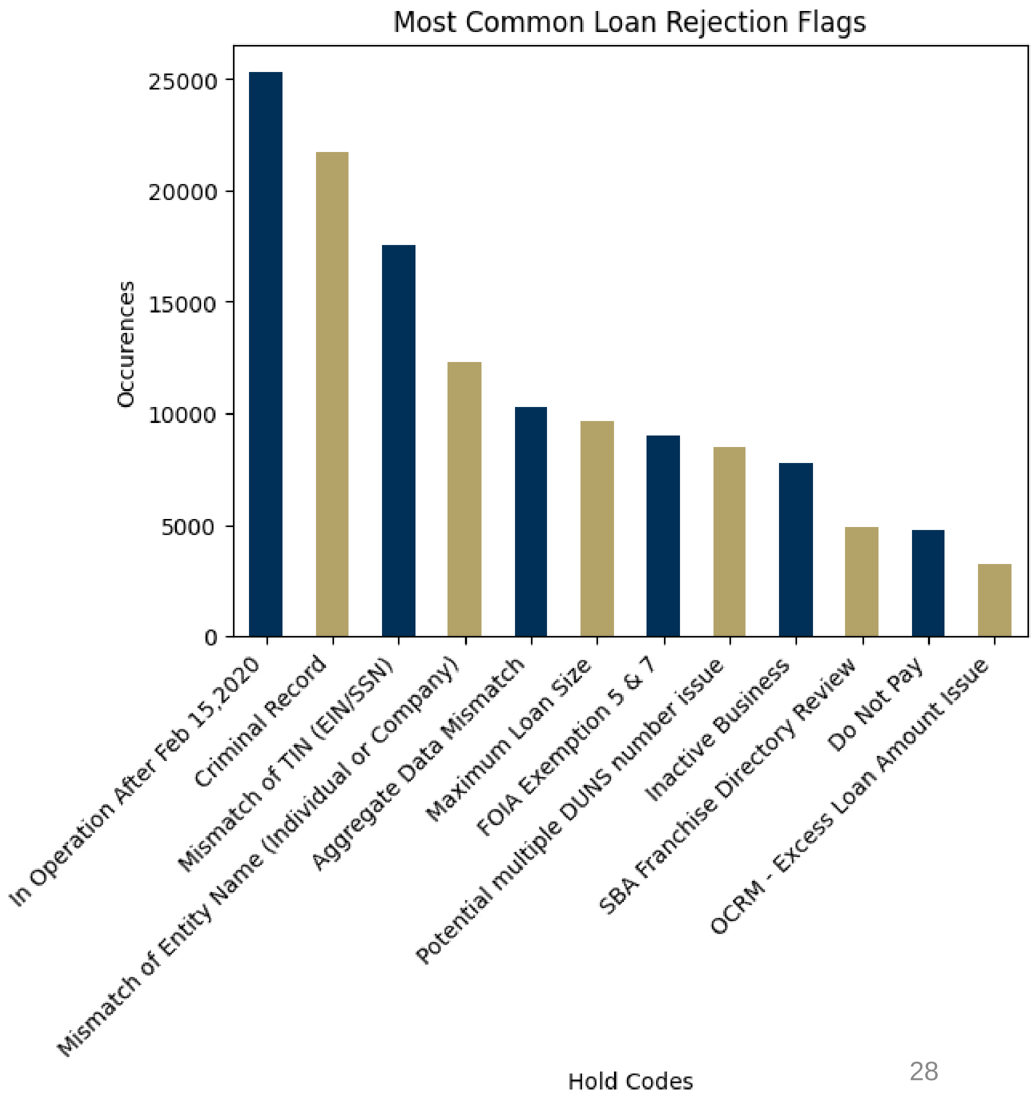
- Data available was for rejected applications, not fraudulent loans
- We are unable to complete this investigation, due to data not being available for fraudulent loans

Suspect Loan Analysis: EDA

25.27%
of applications rejected

12
flags on
most-rejected applications





Suspect Loan Analysis: Conclusions

Our Goal	Major Challenges	Our Action
Originally, our goal was to develop a fraudulent loan detection model based on application characteristics using a dataset of PPP loans flagged for review	• Lack of matching unique identifiers between SBA PPP loan applicant data and rejected application data • No alternative unique identifier combination to join SBA PPP datasets • Lack of detailed data dictionary or many features for suspect loan data • Missing hold codes (government only released 43 of the possible hold codes in the dataset)	• We conducted a literature review and found research that indicated that this analysis was possible • However, we found that the only available data was for rejected applications, not loans that were accepted and later flagged • After analysis of the available data we determined that we would remove this goal

Recovery Analysis: Modeling Approach

FEATURE ENGINEERING

Develop secondary variables

- Response variable:
  - Recovery rate
- Added explanatory variables:
  - Percent of jobs covered by loans
  - Loan dollars per job
  - COVID case rate, death rate
  - Population density
  - Total population

FEATURE SELECTION

Remove highly collinear variables and select most-correlated features

We tested several feature selection methods, including:

- Ridge regression
- Lasso regression
- Auto-selection in random forest and XGBoost models

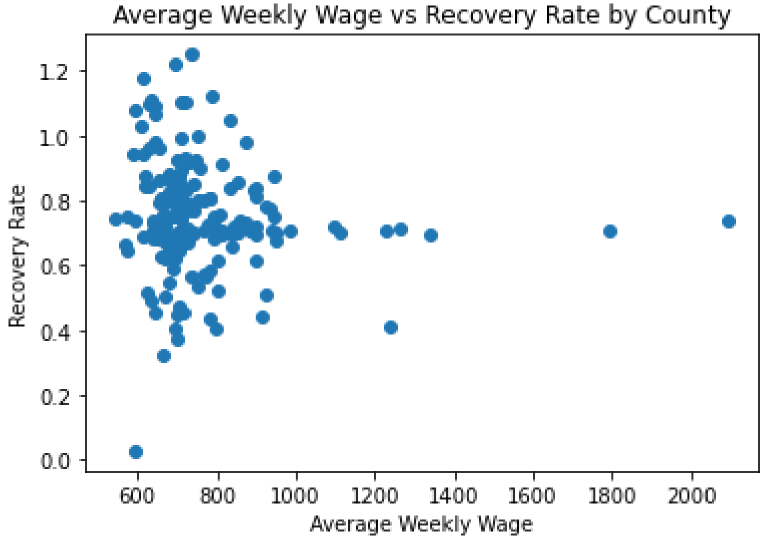
MODELING

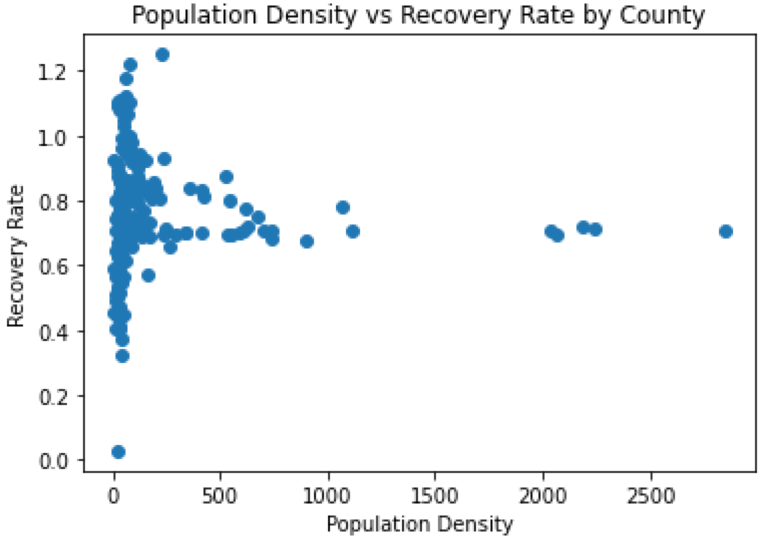
Test linear regression approaches to determine whether PPP loans are correlated with recovery

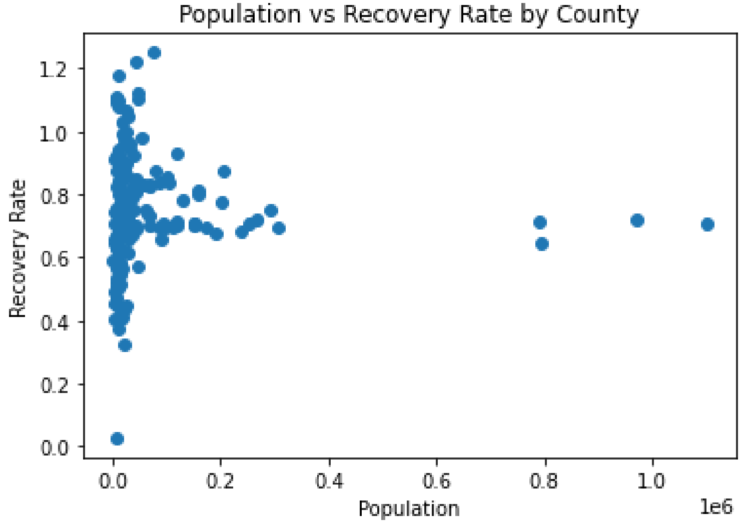
Models tested:

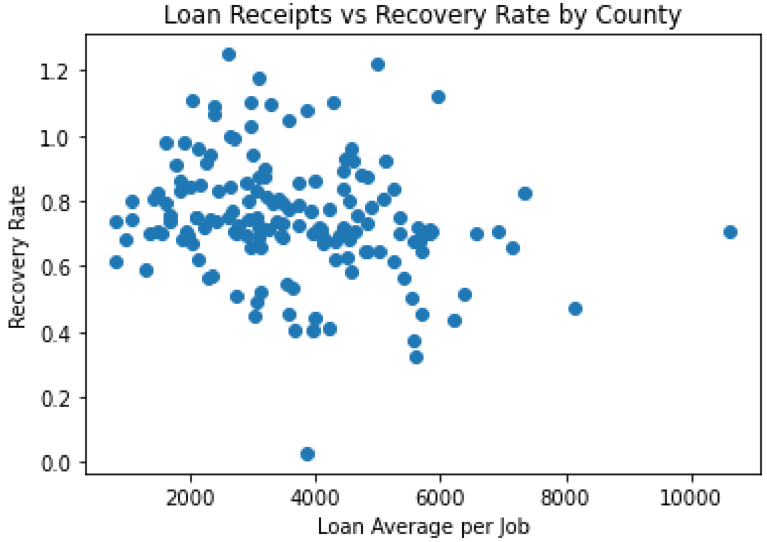
- Multivariate linear regression
- Lasso regression
- Random forest regression
- Extreme gradient boosting model

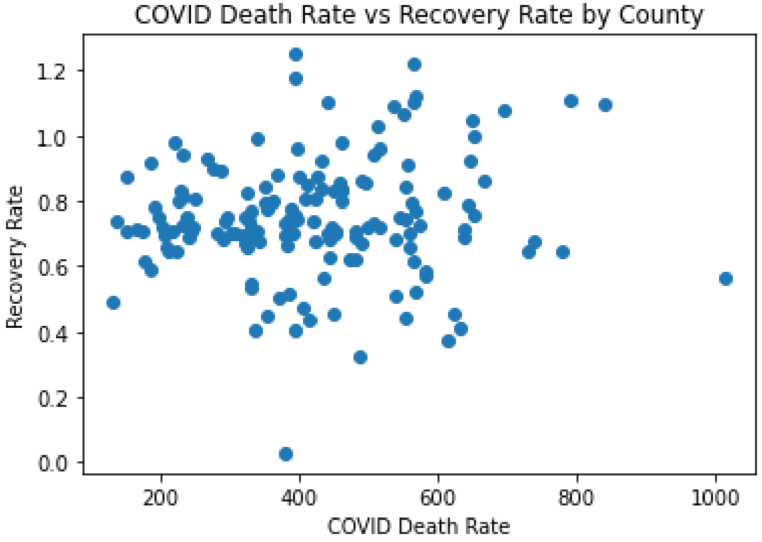
Recovery Analysis: Correlations

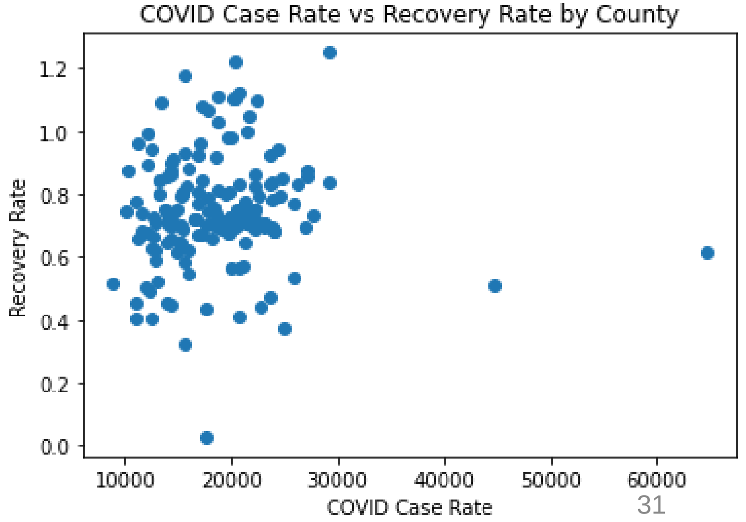












Recovery Analysis: Results

Model	R-Squared	MAPE
Multiple Linear Regression	0.469047	1.2116
Lasso Regression	-0.229622	1.30104
Random Forest Regression	0.227473	0.980833
XGBoost	-0.0576307	1.26111

Recovery Analysis: Conclusions

- When looking at variables that intuitively may impact recovery on an individual basis, none appeared to be particularly correlated with recovery rate.
- All attempted regression modeling techniques yielded models with poor linear fit and large errors when predicting on the test data set.
- We did not pursue further analysis on recovery rate given the modeling limitations present with available data.

Model	R-Squared	MAPE
Multiple Linear Regression	0.469047	1.2116
Lasso Regression	-0.229622	1.30104
Random Forest Regression	0.227473	0.980833
XGBoost	-0.0576307	1.26111