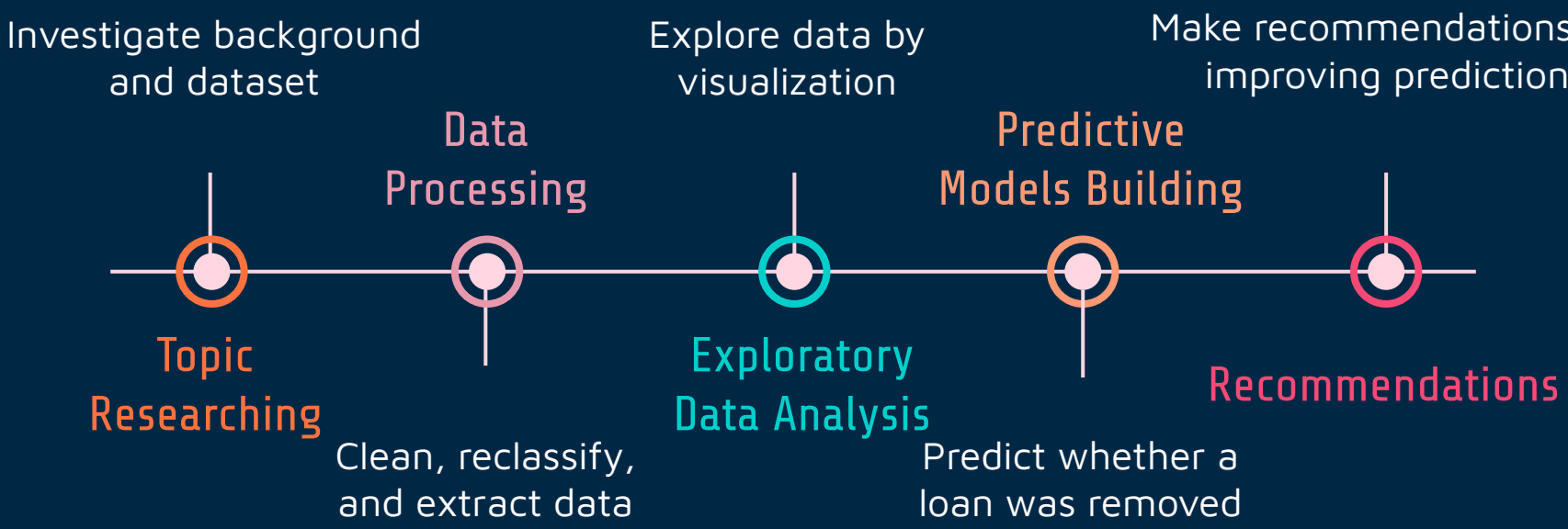
Predicting Removed Loan – Paycheck Protection Program Application

U.S. Small Business Administration

Team 4

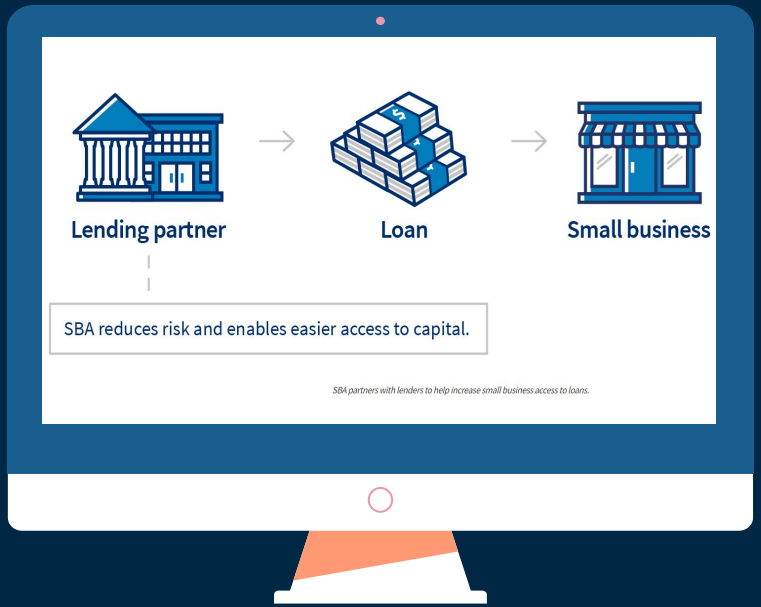
Chia-Lin Tsai
Wang-Han Li
Chung-Hao Lee

OUR PROCESS


Investigate background and dataset
Topic Researching
Data Processing
Clean, reclassify, and extract data
Explore data by visualization
Exploratory Data Analysis
Predictive Models Building
Predict whether a loan was removed
Make recommendations for improving prediction
Recommendations

Introduction

- Small Business Administration (SBA) The only cabinet-level federal agency fully dedicated to **small business**.
- Paycheck Protection Program (PPP) A **$900-billion-budget** SBA-backed **loan** that helps small businesses keep their workforce employed during the **COVID-19 crisis**.
- The SBA has regularly released data on approved applications, but it also has **removed** some **previous applications**.



DATASET

Two Datasets, describing loans to businesses in **Georgia**.

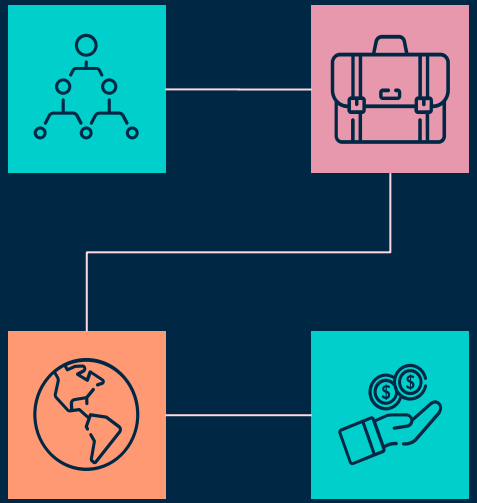
- More than 25,000 loans to GA businesses that were **removed** from the PPP database.
- Around 550,000 loans to GA businesses that were **non-removed** in the PPP database.
- There are **41 variables**, including borrowers and lenders' information, such as address, business type, loan status.
- The data period is from 2020 to 2021.

DATA Processing

Business Type

Reclassify to four types

- Non-profit
- Sole company
- Corporation
- Others



NAICS Code

Extract the first two digits and replace N/A with “81”

* 81: Other Services

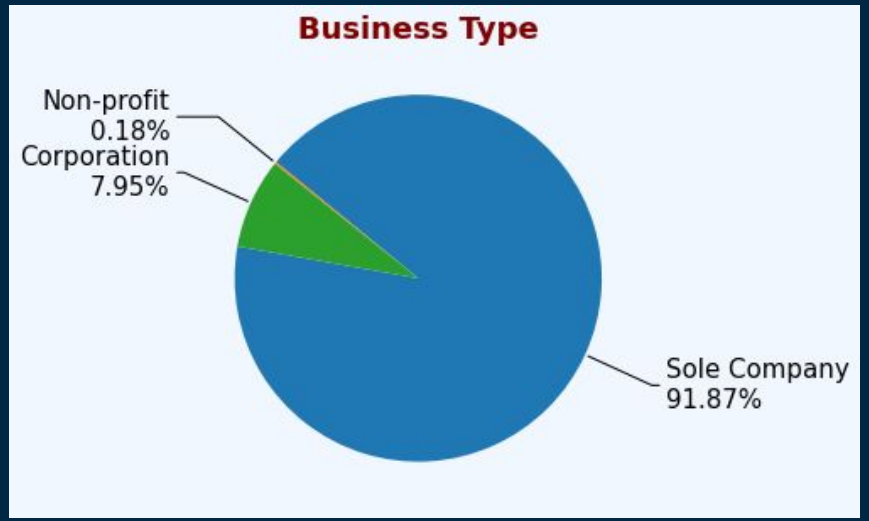
ZIP Code

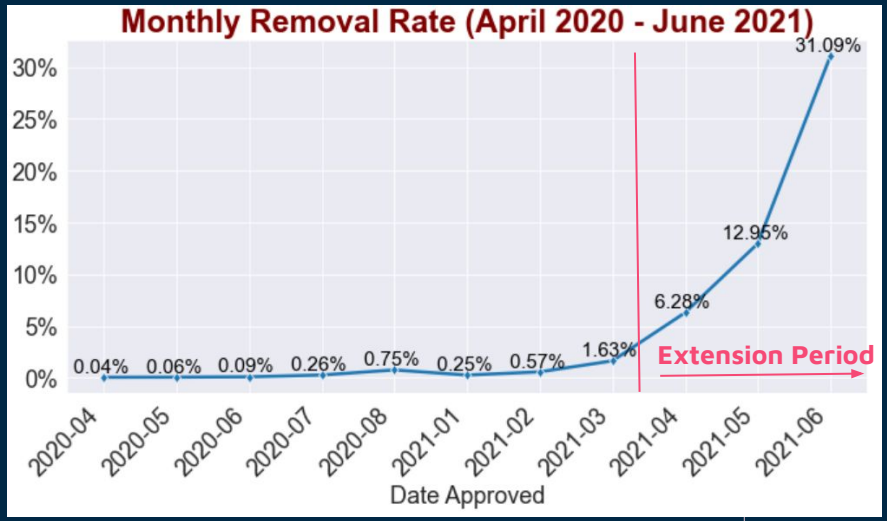
Extract the first five digits of ZIP code

Forgiveness Amount

Replace N/A with 0

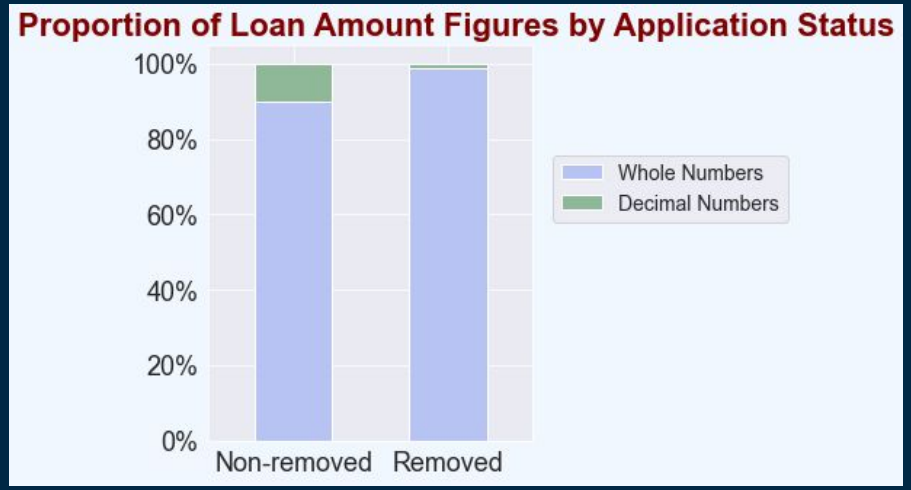
Characteristics of Removed Applications

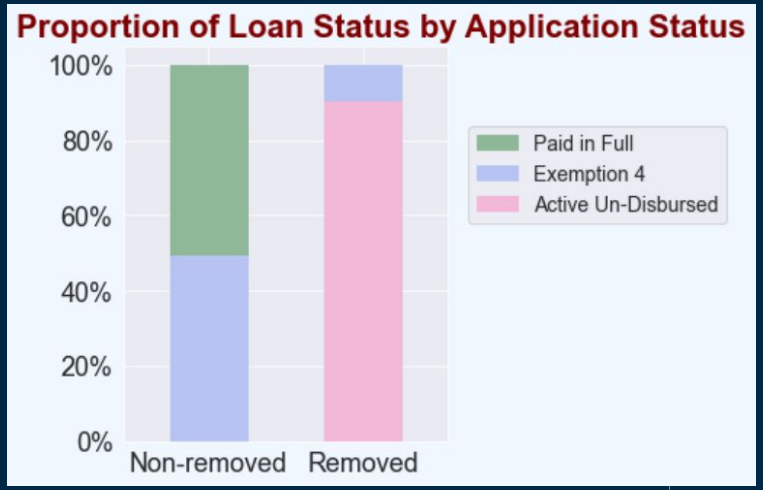




- **Sole company** accounted for more than **90%** of the removed applications.
- The removal rate of application soared **after March 2021**, skyrocketing to **31.09%** in June.

Comparison of Removed and Non-removed





- **Removed** applications had a **lower percentage of initial approval amount** with **decimals** compared to non-removed applications.

- In the **removed** applications, **active un-disbursed** accounted for nearly **90%**. While in the non-removed dataset, applications are almost equally divided between paid in full and exemption 4.

Models to predict whether a loan was removed

Input Variables :

Naics_code

Loan_status

Lmi_indicator

Hubzone_indicator

Business_age_description

Business_type_classification

Yearmonth

Amount_diff

If_decimal_equal_zero

Amount

Jobs_retained

Forgiveness_amount

Borrower_lat

Borrower_lng

Servicing_lender_lat

Servicing_lender_lng

Derivative variables

Convert to Dummies

Convert zip code to latitude and longitude

Predictive Models: XGBoost and Logistic Regression

Split:
Training 70%
Testing 30%

XGBoost

Accuracy : 0.995
AUC : 0.949
Sensitivity : 0.899

Actual \ Predicted	r	nr
r	6854	765
nr	92	165005

Logistic Regression

Accuracy : 0.975
AUC : 0.765
Sensitivity : 0.534

Actual \ Predicted	r	nr
r	4068	3551
nr	700	164397

AUC: Probability that the classifier will be able to distinguish between classes
Sensitivity: Proportion of actual positive cases which got predicted as positive

Recommendations

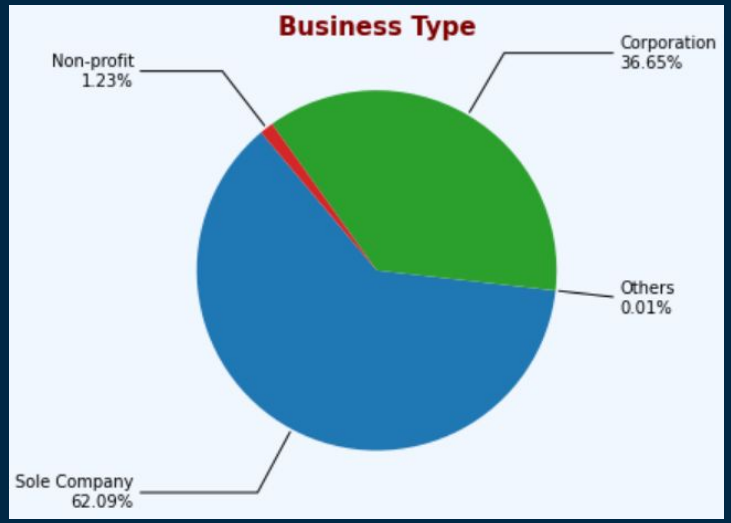
More research can do to find out removed loans:

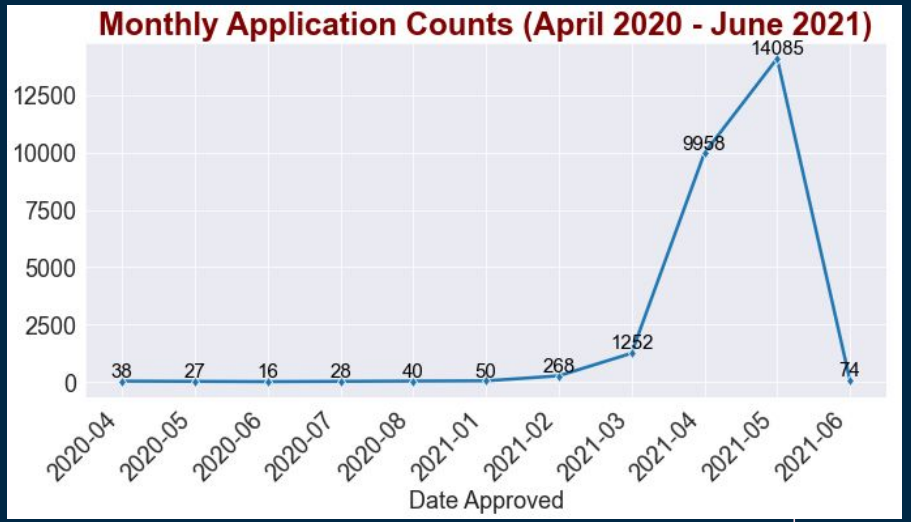
- Non-registered businesses
- Multiple loans at a residential address
  - Increase **reliability** of a loan borrower

- Submission date
  - Judge if an application is submitted in the **extension period**

Thank You

Supplementary





XGBoost

Extreme Gradient Boosting (XGBoost) iteratively train an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model. The final prediction is a weighted sum of all of the tree predictions.