TEAM 30

PREDICTION & ANALYSIS OF REMOVED LOAN APPLICATIONS

PAYCHECK PROTECTION PROGRAM

UPASANA MOHAPATRA
LING FANG
YU-TUNG CHANG

Agenda

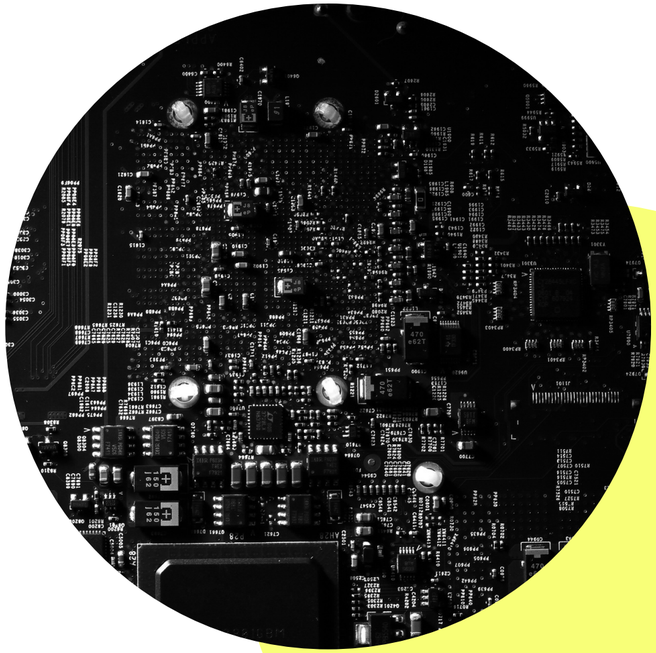
1. Introduction
2. Data Exploratory Analysis
3. Detailed Analysis on features
4. Hypothesis
5. Models Selection & Training
6. Key Insights
7. Limitation & Future Work

Paycheck Protection Program

U.S. Small Business Administration (SBA)

- Enacted by Congress in 2020 to respond to the economic impact of the COVID-19 pandemic
- Loans worth $800 billion provided to small businesses
- The loans are administrated by private lenders
- 11.5 million approved applications released with little to no information on cause of removal of some of the loans
- Focusing area of analysis: State of **Georgia**

Data Exploratory Analysis



Q1

Why were the loans removed?

Q2

Differentiating factors between datasets

Q3

Predict which loans will be further removed

Key Variables

Borrower Details

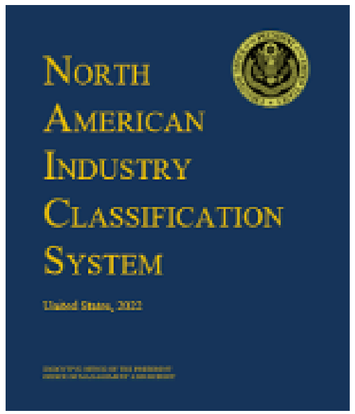
- Loan Amount
- Congressional District
- City
- Loan Status
- Service Industry(derived from NAICS)
- Jobs Retained
- Business Type
- Undisbursed Amount

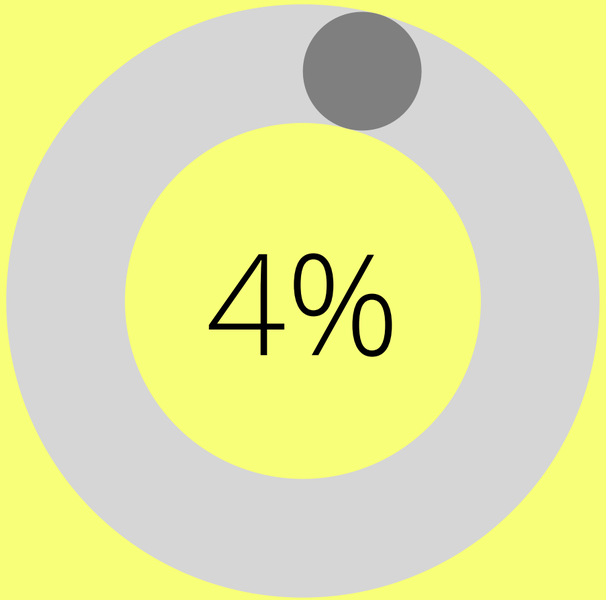
Lender Details

- Lender
- Lender Address

Additional Datasets

United States Census Bureau





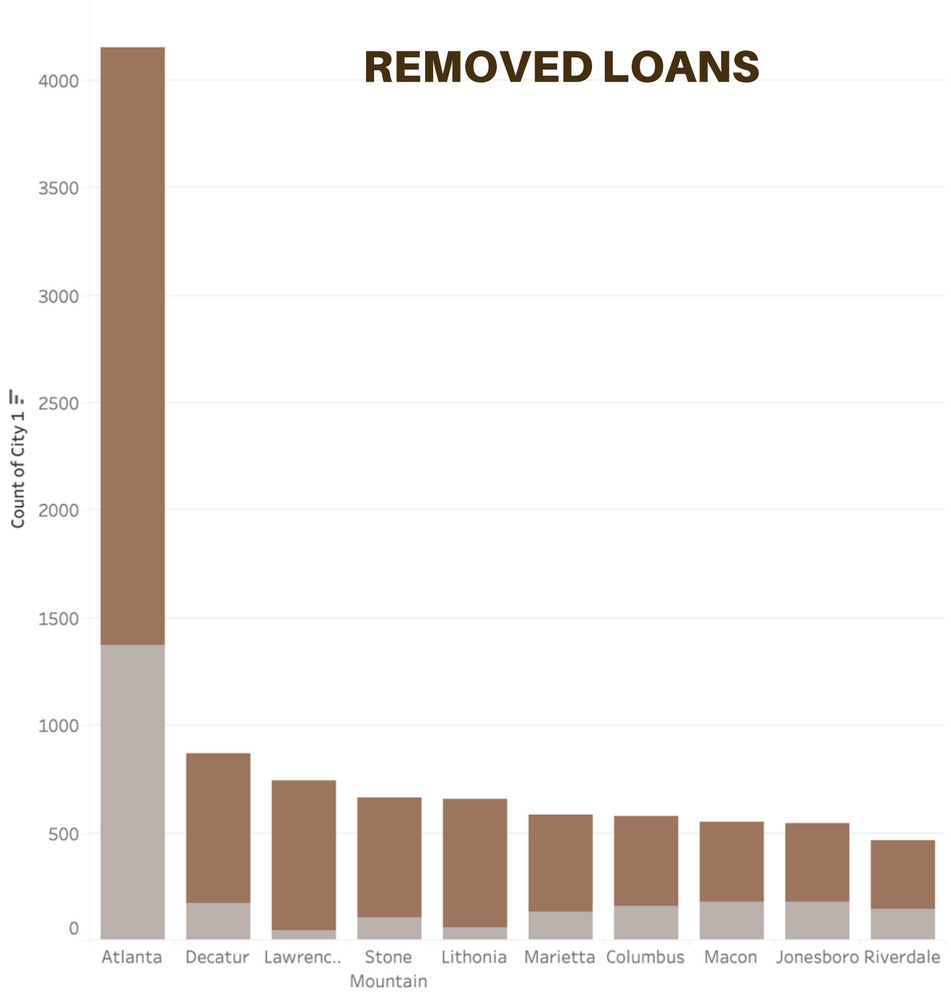
Removed dataset is 4% of the size of the full dataset

The average maximum loan approval amount for the public dataset is >2X compared to the removed loans

Top Cities

Atlanta, Decatur, Lawrenceville, Stone Mountain, Lithonia

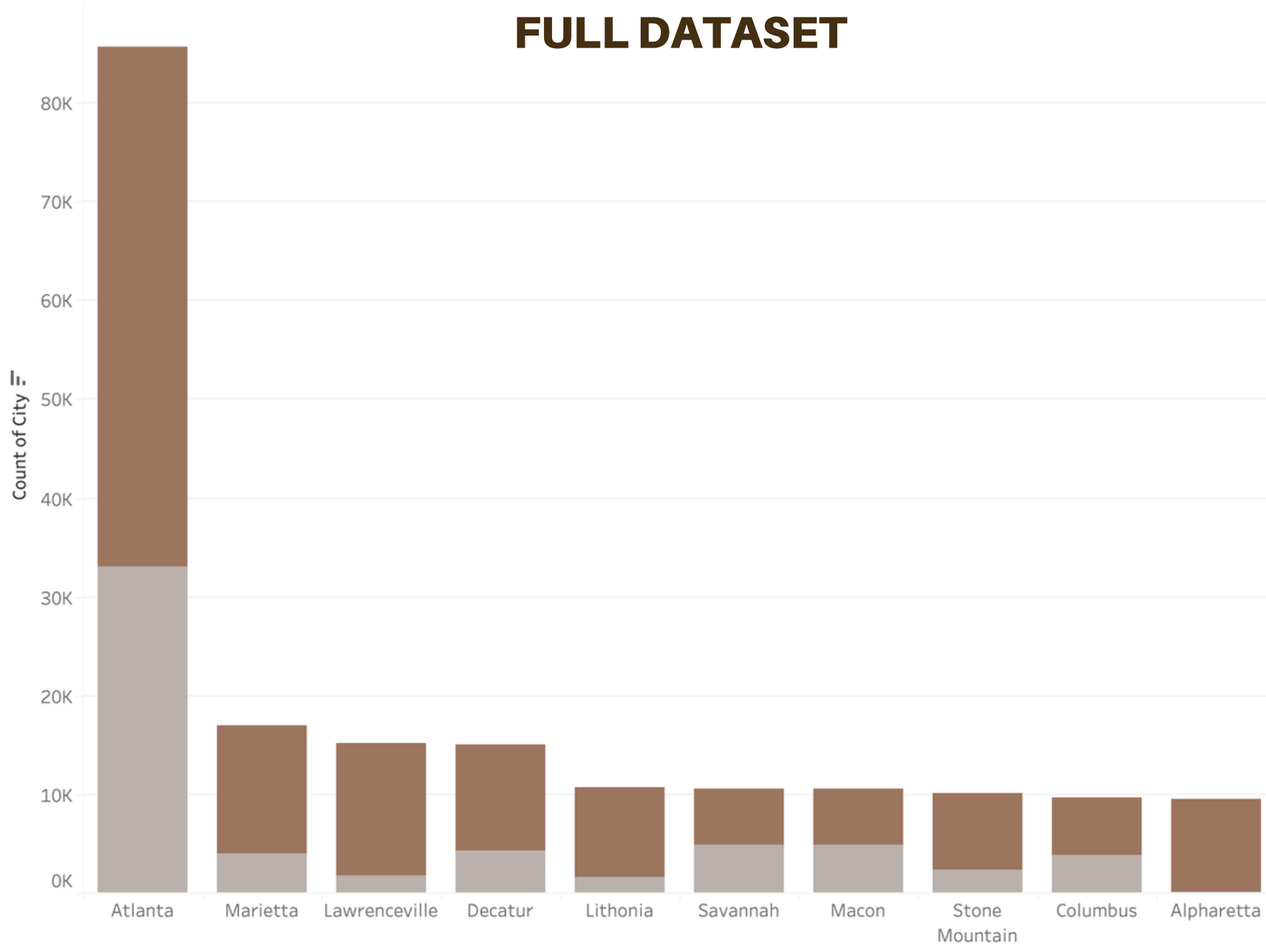
REMOVED LOANS



Count of City 1 for each City 1. Color shows details about Hubzone Indicator. The view is filtered on City 1, which keeps 10 of 517 members.

Atlanta, Marieta, Lawrenceville, Decatur, Lithonia

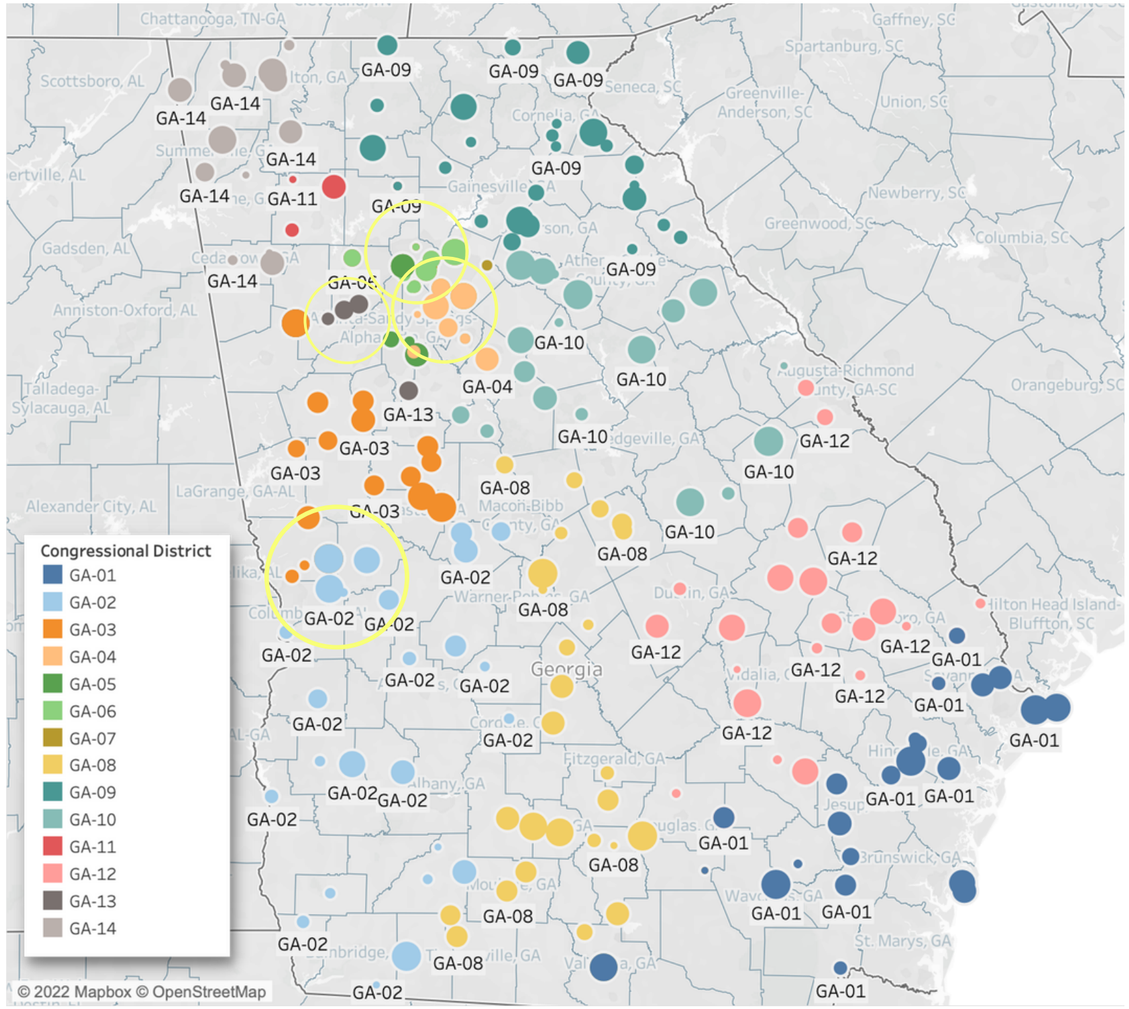
FULL DATASET



Count of City 1 for each City 1. Color shows details about Hubzone Indicator. The view is filtered on City 1, which has multiple members selected.

Top Districts

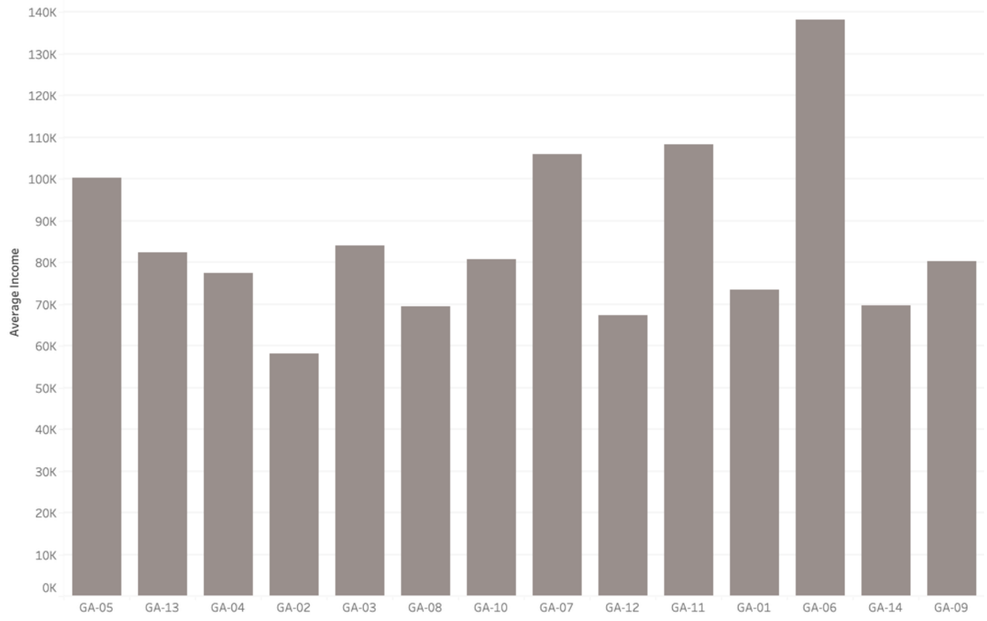
DISTRICT



Top Districts - Removed

TOTAL AMOUNT ACROSS DISTRICT & INDSUTRY

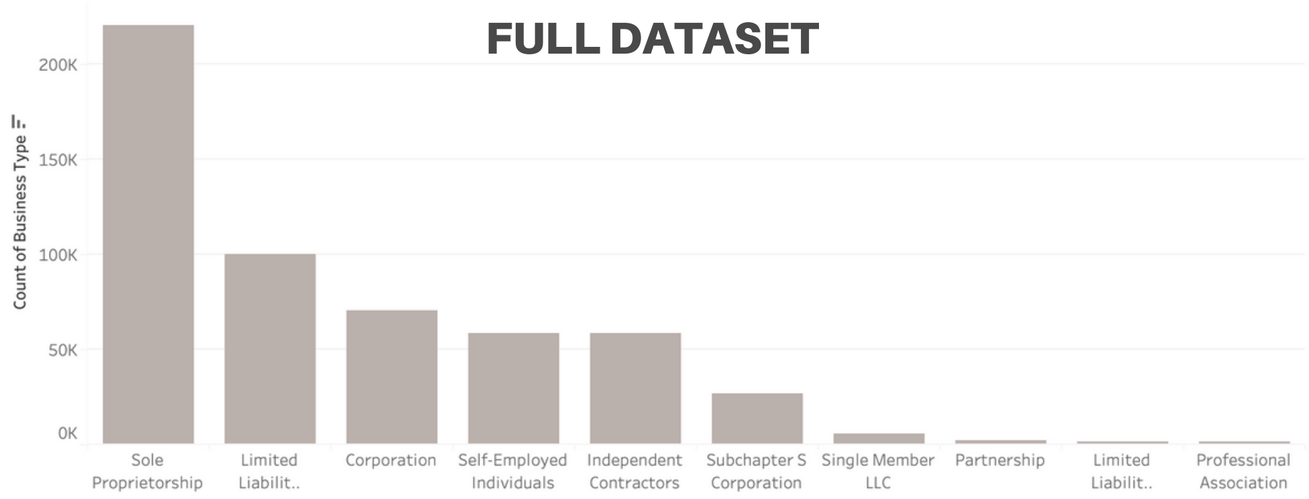
	GA-05	GA-13	GA-04	GA-02	GA-03	GA-08	GA-10	GA-07	GA-12	GA-11	GA-01	GA-06	GA-14	GA-09
Personal and Laundry Services	1,051	954	832	551	481	375	334	344	273	228	242	187	105	87
Professional, Scientific, and Technical Services	388	326	314	130	129	81	94	177	61	127	64	132	45	43
Administrative and Support Services	337	378	295	213	264	161	131	130	87	103	108	84	66	47
Food Services and Drinking Places	192	208	179	81	78	59	65	65	52	73	53	54	41	16
Transit and Ground Passenger Transportation	178	165	181	52	71	44	45	104	50	59	30	52	17	15
Performing Arts, Spectator Sports, and Relate..	172	178	125	57	41	38	33	56	40	56	28	53	19	10
Construction of Buildings	143	190	165	69	80	47	60	66	39	52	26	32	32	28
Clothing and Clothing Accessories Stores	141	128	116	71	74	69	29	47	48	29	28	27	12	8
Truck Transportation	111	163	191	84	77	63	65	76	54	48	73	30	14	20
Health and Personal Care Stores	63	38	58	62	37	17	26	24	23	20	13	16	15	7

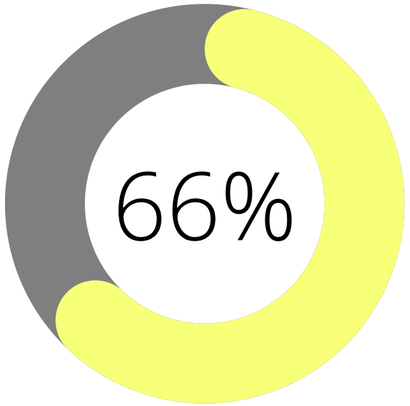


Lower average incomes could indicate higher demands for loans

Key Figures - Comparison

BUSINESS TYPE





66% of the business type of removed loans is **Sole Proprietorship**
(compared to 40% from full dataset)

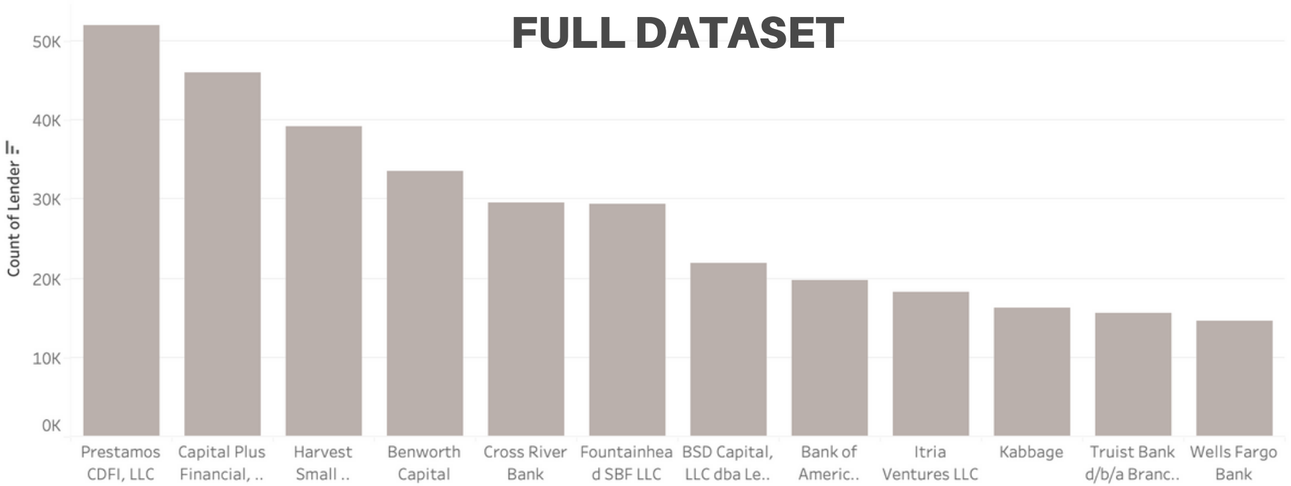
Key Figures - Comparison

LENDER

REMOVED LOANS

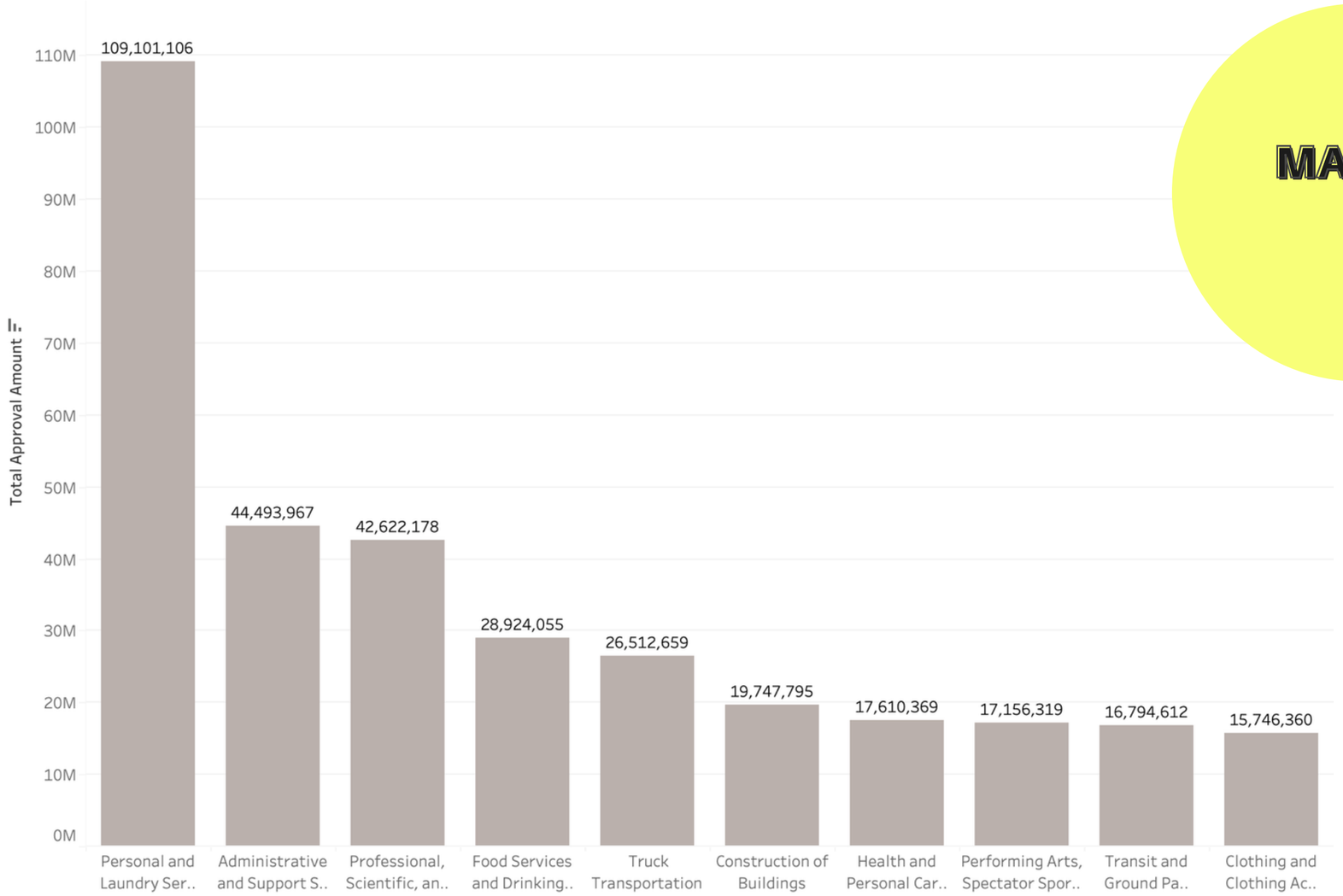
- Capital Plus Financial
- Prestamos CDFI
- BSD Capital
- Customers Bank
- Leader Bank

FULL DATASET



- Prestamos CDFI
- Capital Plus Financial
- Harvest Small Business Finanace
- Benworth Capital
- Cross River Bank

Top 5 Industries - Removed Loan Applications



Sum of Amount for each Naics Code 3. The marks are labeled by sum of Amount. The view is filtered on Naics Code 3, which keeps 10 of 94 members.

MAJORITY OF THE REMOVED LOANS WERE FROM B2C INDUSTRIES

- Personal and Laundry Services
  (Beauty Salons, Barber Shops)
- Administrative and Support Services
- Professional, Scientific, and Technical Services
- Food Services and Drinking Places
- Truck Transportation

Key Hypothesis

Now that we've seen the data, let's build a few hypothesis we can use to test our analysis

1

B2C businesses in the full dataset are likely to get removed

Avg loan amount for full dataset is >2X of the removed loans

2

Top Cities with max no. of removed loans might be removed from the full dataset

3

Top lenders with maximum removed loans will likely be removed from the full dataset

4

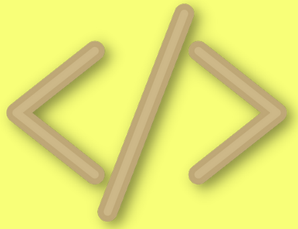
Loans with Exemption Status 4 will likely be removed from the full dataset

5

Loans given to business type - Self-Proprietorship, Independent Contractors, Self Employed Individuals will likely be removed from the dataset

Model Training

Choice of model and implementation



Neural Network Binary Classification Model

Training data: 75% removed loans and 25% full dataset loans

Data Cleaning

- **Remove NA values**
- **Get categorial variables**
- **Replace missing value**

One-hot encoding to create dummies for features that add values to model

Reducing Mean Average Error for the model and predicting its accuracy

Predictions on the entire sample of unknown data to predict removal of loans

Code Snippets

```
[ ]  1 # taking 75 precent from removed and 25 from full for training
     2
     3 np.random.seed(100)
     4 removed_training = removed_label.sample(frac = 0.75, replace = False,random_state =100)
     5 full_training = full_label_sampled.sample(frac = 0.25, replace = False,random_state =100)
     6 training_data = pd.concat([removed_training,full_training ], axis=0)
     7
```

```
[ ]  1 #dropping null values
     2 training_data.dropna(subset = ["amount","city","naics_code","business_type","jobs_retained","lender","congressional_district","loan_status","initial_approval_amount","current_approval_amount","undisbu
     3
```

```
[ ]  1 #divide labels and training data
     2 X = training_data.drop("removal_status", axis=1)
     3 y = training_data["removal_status"]
```

```
[ ]  1 ### now splitting total training data for training
     2 from sklearn.model_selection import train_test_split
     3 X_train, X_test, y_train, y_test = train_test_split(X,
     4                                                     y,
     5                                                     test_size=0.2,
     6                                                     random_state=42)
```

```
[ ]  1 # convert data for NN to understand (one hot encoding)
     2 from sklearn.compose import make_column_transformer
     3 from sklearn.preprocessing import MinMaxScaler, OneHotEncoder
     4
     5
     6
     7 ct = make_column_transformer( # get all values between 0 and 1
     8     (MinMaxScaler(), ["amount","initial_approval_amount","current_approval_amount","undisbursed_amount", "jobs_retained"]),
     9     (OneHotEncoder(handle_unknown="ignore"), ["city","naics_code", "business_type","lender","congressional_district","loan_status","servicing_lender_city","servicing_lender_state","hubzone_indicator",
    10 )
    11 ct.fit(new)
```

```
1 #model creation
2 import tensorflow as tf
3
4 tf.random.set_seed(100)
5
6 # Add an extra layer and increase number of units
7 model_1 = tf.keras.Sequential([ # 100 units
8   tf.keras.layers.Dense(1000), # 10 units
9   tf.keras.layers.Dense(100),
10  tf.keras.layers.Dense(10),
11  tf.keras.layers.Dense(1) # 1 unit (important for output layer)
12 ])
13
14 # Compile the model
15 model_1.compile(loss=tf.keras.losses.mae,
16                           optimizer=tf.keras.optimizers.Adam(), # Adam works but SGD doesn't
17                           metrics=['mae'])
18
19 # Fit the model and save the history
20 history = model_1.fit(X_train_normal, y_train, epochs=5, verbose=1)
```

```
Epoch 1/5
646/646 [==============================] - 97s 148ms/step - loss: 0.1613 - mae: 0.1613
Epoch 2/5
646/646 [==============================] - 95s 147ms/step - loss: 0.1360 - mae: 0.1360
Epoch 3/5
646/646 [==============================] - 95s 147ms/step - loss: 0.1108 - mae: 0.1108
Epoch 4/5
646/646 [==============================] - 95s 147ms/step - loss: 0.1037 - mae: 0.1037
Epoch 5/5
646/646 [==============================] - 96s 148ms/step - loss: 0.0971 - mae: 0.0971
```

```
[ ]  1 #model evaluation on test data
     2 model_1.evaluate(X_test_normal, y_test)
```

```
162/162 [==============================] - 2s 10ms/step - loss: 0.1333 - mae: 0.1333
[0.1332939863204956, 0.1332939863204956]
```

```
[ ]  1 #single prediction check for model on test data
     2 model_1.predict(X_test_normal[0].reshape(1,24611))
```

```
array([[0.00469012]], dtype=float32)
```

```
[ ]  1 ct.fit(new_sampled_df)
```

```
[ ]  1 ## sample data set transform
     2 ## transform the full data for predictions
     3 full_predict_data = ct.transform(new_sampled_df)
```

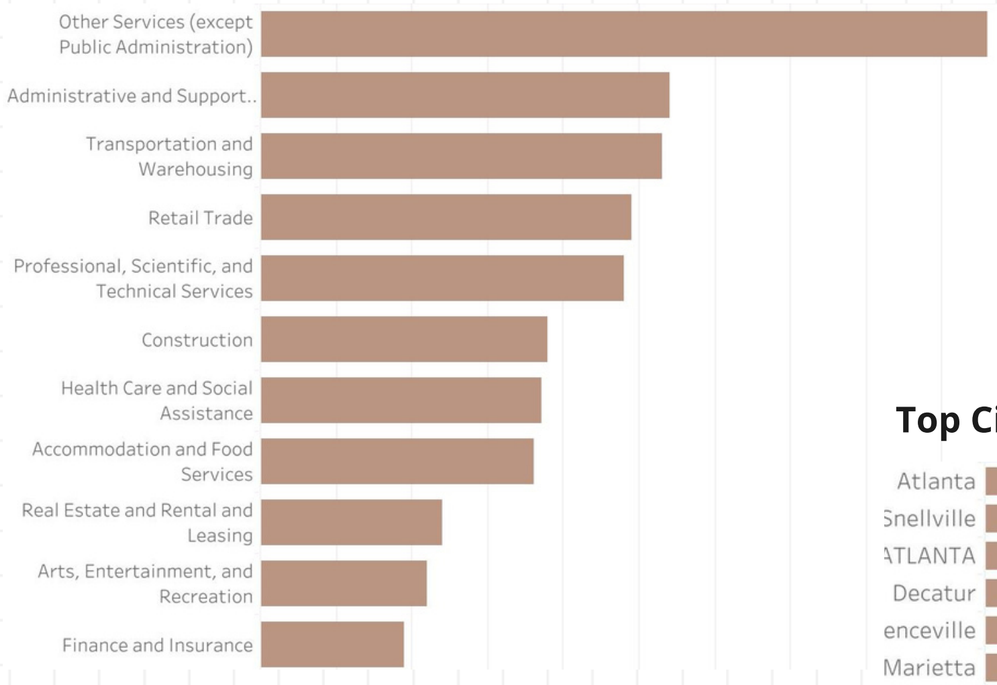
```
1 print(full_predict_data.shape[0])
2 new_sampled_df["status"] =0
3 df_final = pd.DataFrame(columns = ["amount","city","naics_code","business_type","jobs_retained","lender","congressional_district","loan_status","initial_approval_amount","current_approval_amount","undi
4
5 for i in range(0,full_predict_data.shape[0]):
6 a = model_1.predict(full_predict_data[i])
7
8 if(0.75<a):
9   new_sampled_df.at[i, "status"] = 1
10 df_final.loc[i] = new_sampled_df.loc[i]
11 df_final.loc[i] = new_sampled_df.loc[i]
12 print(i)
13
14 df_final.to_csv("filtered_full_dataset", sep='\t', encoding='utf-8')
15
```

0.09% of the entire sample predicted to be removed

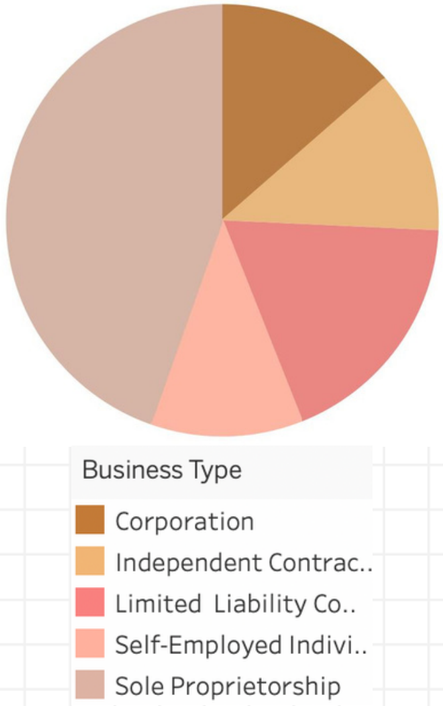
from the full data set

542 loans identified out of ~555,000

Industries predicted to be removed based on count of applications



Based on count of business type

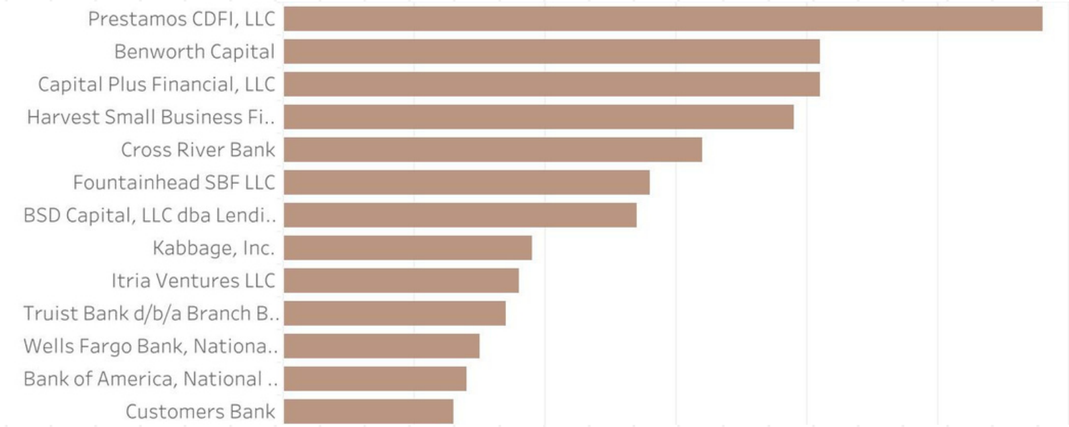


Top Cities to be removed based on count of applications

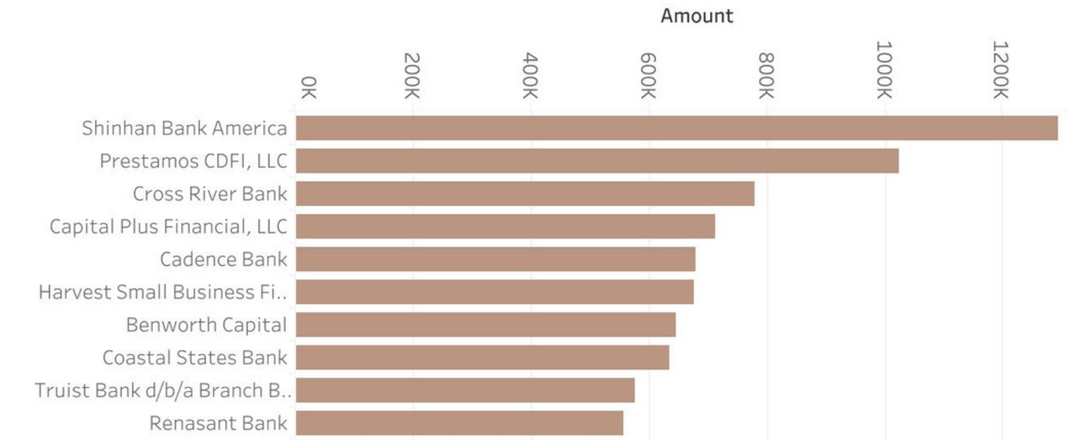
Few more observations

Lenders, Loan Status Analysis

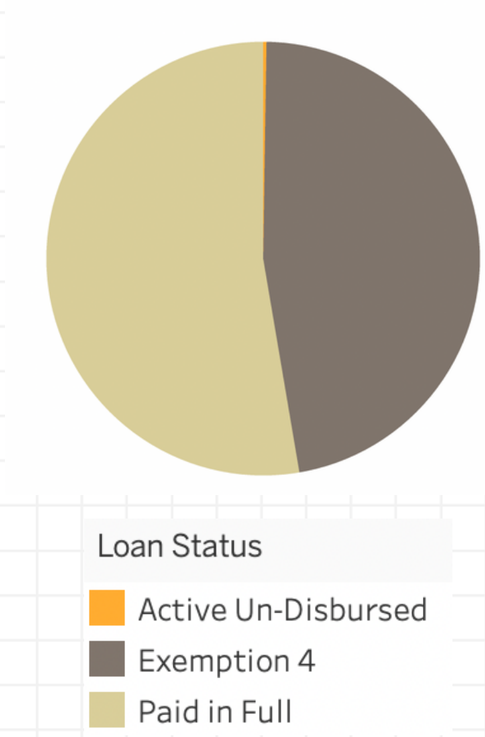
Based on count of loans per lender



Based on sum of loan amount per lender



Based on count of loans by loan status



So basically...

This worked

B2C businesses in the full dataset are likely to get removed

Miscellaneous Service Industries, Retail, Administrative Services are predicted to be removed

Top Cities with max no. of removed loans might be removed from the full dataset

Atlanta, Decantur, Lawrenceville, Ethonia, Stone Mountain are some common cities

Top lenders with maximum removed loans will likely be removed from the full dataset

Capital Plus, Prestamos , Bensworth Capital, Customers Banks

Loans with Exemption Status 4 will likely be removed from the full dataset

~45% of the data being removed: Exemption Status 4
53% of the data being removed: Paid In Full

Loans given to business type - Self-Proprietorship, Independent Contractors, Self Employed Individuals will likely be removed from the dataset

Sole Proprietors, Independent Contractors with some difference from removed dataset with Corporations, LLC

What's Next?

LIMITATIONS AND FURTHER STUDY

Loan Status in the removed dataset could be biased

- The full dataset consists of a lot of loans that have been paid in full,
- Training data had no information on these.
- *Our model accurately predicted the majority of the loans that were fully paid to be removed*

Additional Analysis

Borrower Information

- Company Founding Year
- Loan History Record
- Financial Balance Sheets

Lender Information

- The size of lender
- Capital Scale

More Information

- Economic Factor
- COVID-19 cases

Thank you!

We are a group of MS in Marketing Analytics Students who have very recently stepped into the world of data analytics(with limited knowledge on the subject)! This was the most exciting project we've worked on!

Upasana Mohapatra

upasanam@umd.edu

Ling Fang

lfang423@umd.edu

Yu-Tung Chang

ytc0128@umd.edu