PPP Removed Application

Data Challenge
IC22021

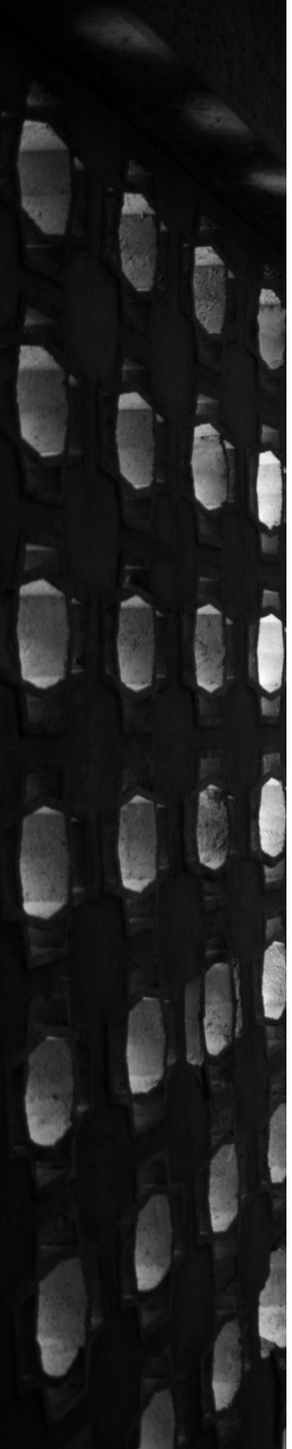
Amola Patel
Amy Chan

Yufei Deng
Ziyu Liu

Paycheck Protection Program is a Small Business Administration backed loan that helps businesses keep their workforce employed during the COVID-19 crisis.

- What are some defining characteristics of the removed loans?
- How do the characteristics of loans that were removed from the PPP data compare to loans that were not removed from the PPP data?
- Is it possible to accurately predict whether or not a loan was removed from the data

- Data Preparation
- Data Cleaning



Workflow

PROPOSE THE HYPOTHESIS

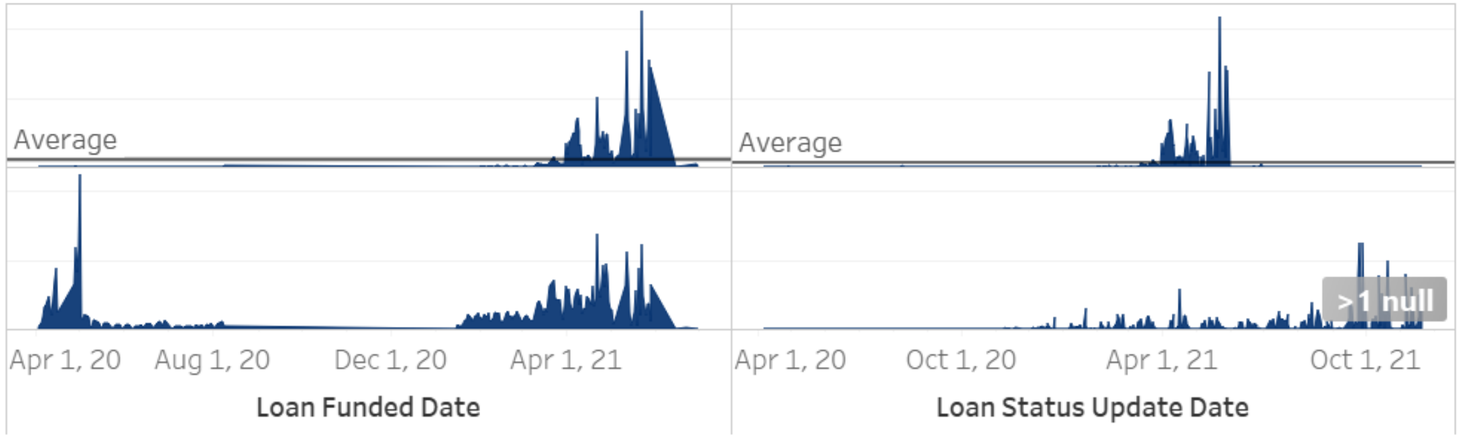
VERIFY THE HYPOTHESIS

STATISTICAL MODELING

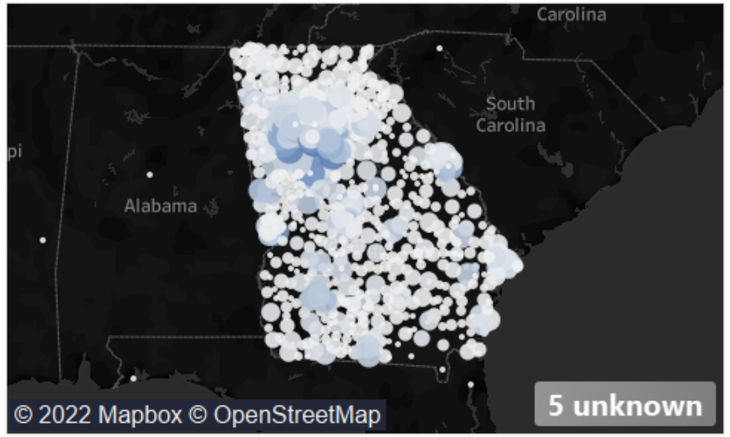
MACHINE LEARNING

Hypothesis

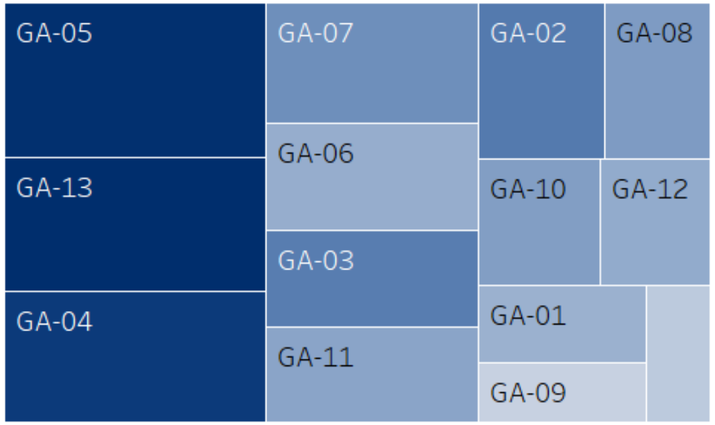
A large number of records updated between April and June were removed from the database



Removed records cluster in a certain area

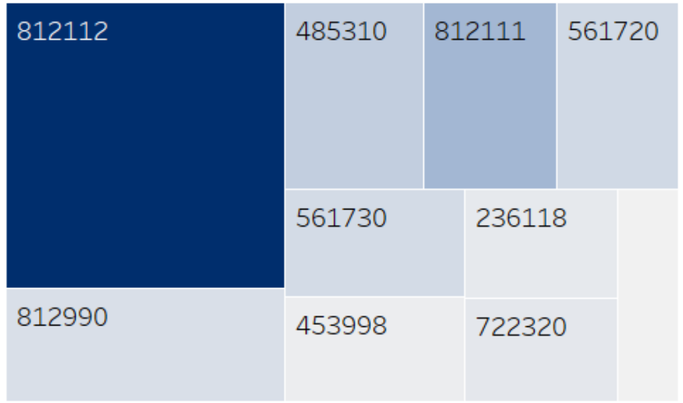


Removed records are mainly from GA-05, GA-13 and GA-04

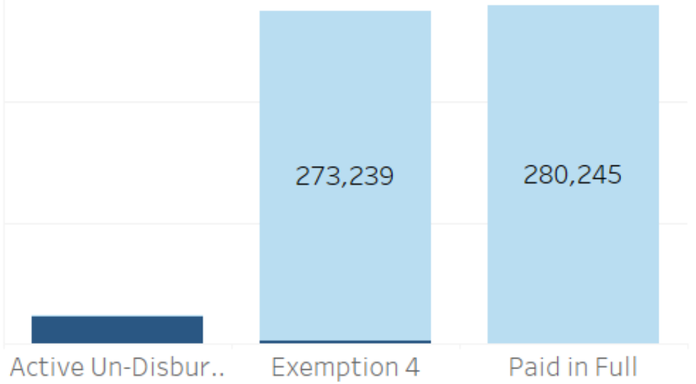


Hypothesis

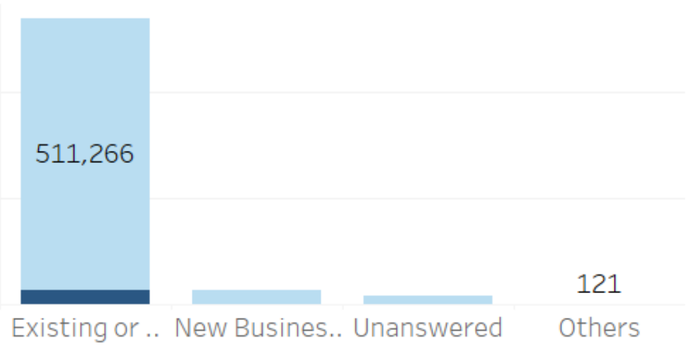
Top industry removed records from is Beauty Salons



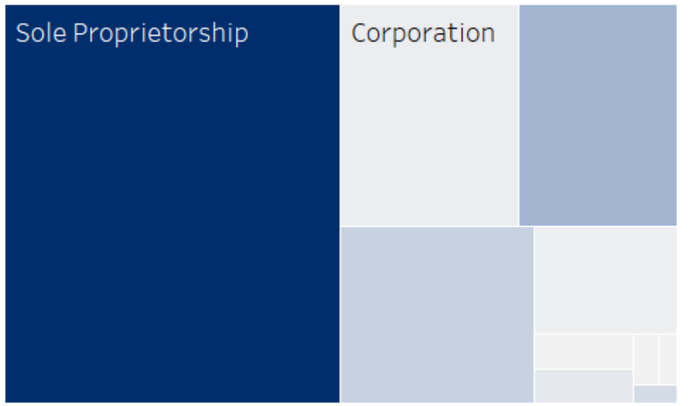
Removed records are mainly in Active Un-Disbursed status



Removed records are only from existing or more than 2 year company

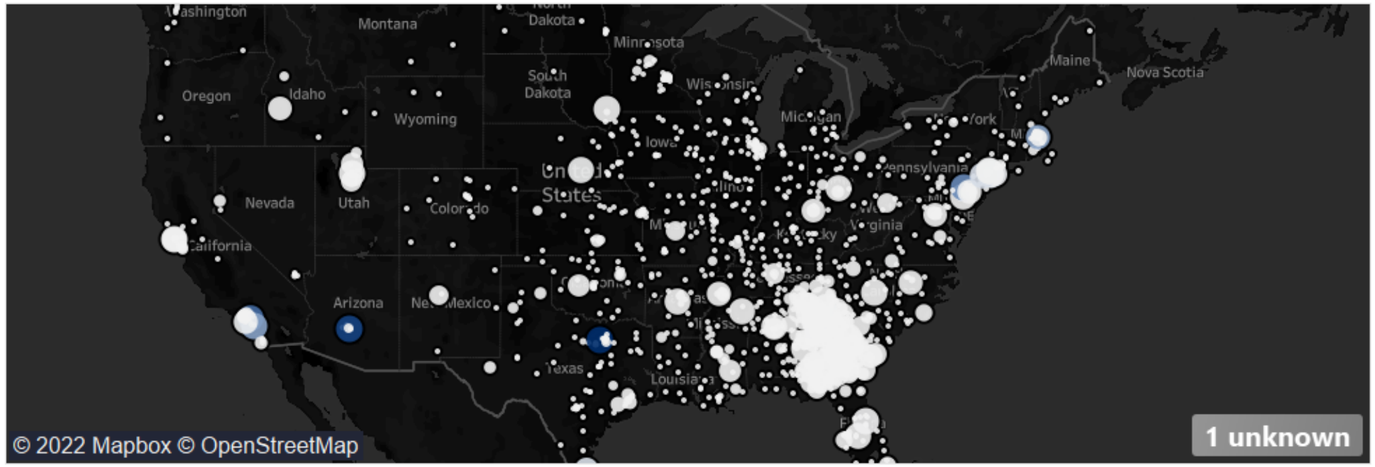


Top business type of removed records is Sole Propietorship

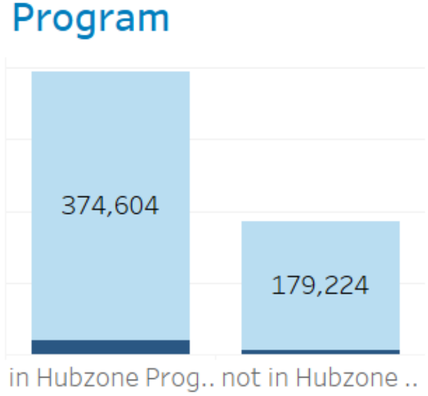


Hypothesis

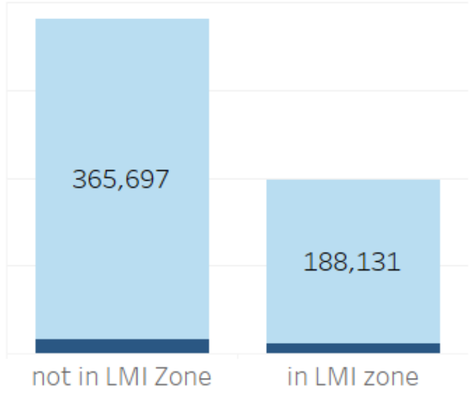
Lender location of removed records is scattered



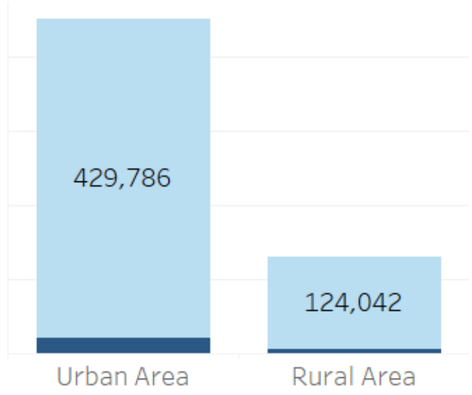
Removed records are mainly in Hubzone Program



Removed records are mainly not in LMI Zone



Removed records are mainly in Urban Area



Statistical Inference

- LOAN STATUS
  - THE ACTIVE UN-DISBURSED STATUS MEANS YOU WERE APPROVED FOR THE PPP LOAN, HAVE SIGNED YOUR CONTRACTS, AND ARE WAITING FOR YOUR FUNDS TO BE DISPURSED.
- STATUS GENERATION CYCLE
  - THE TOTAL NUMBER OF DAYS BETWEEN LOAN FUNDED DATE AND LOAN STATUS DATE

Statistical Modeling

- BINARY LOGISTIC REGRESSION MODEL

Odds Ratio

Status generation cycle

17.1% decrease in odds of being removed for each additional day in length of loan status generation cycle.

Hubzone Program

the odds of being removed for small business companies in Hubzone program is 44% lower than the odds for those not in Hubzone program.

Company Size

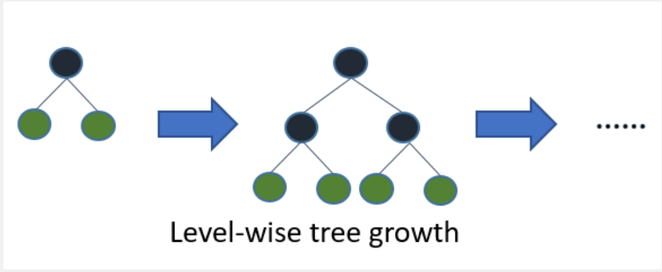
2.6% increase in odds of being removed for each additional employee in small business company.

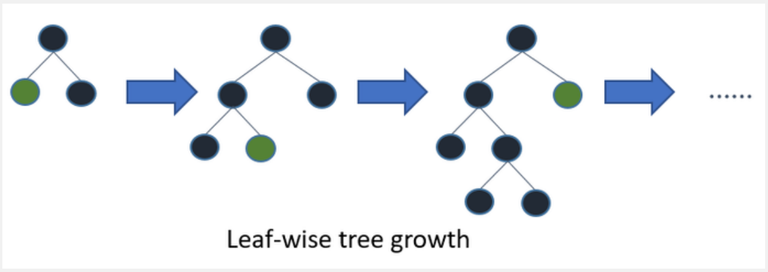
Urban Area

the odds of being removed for small business companies in urban area is 35.3% lower than the odds for those in rural area.

Machine Learning

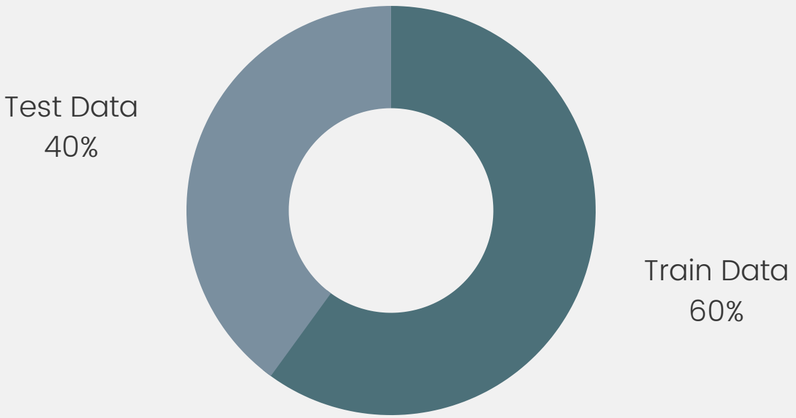
LightGBM Model





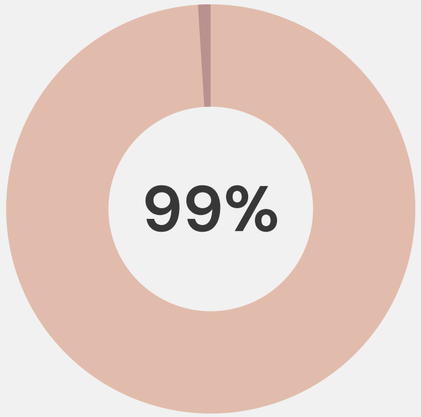
To accurately predict whether or not a loan was removed from the data

- Decision tree algorithm
- Leaf-wise tree growth → more accuracy
- Fast!



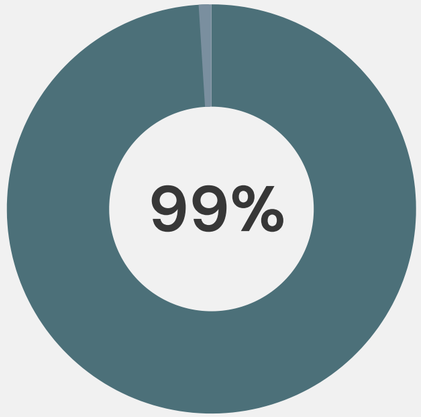
- LightGBM Model training-set accuracy score: 0.9957
- LightGBM Model testing-set accuracy score: 0.9953
- The actual testing and training scores do not differ too vastly.
- This shows that the model is not overfitting.

Metrics



Accuracy

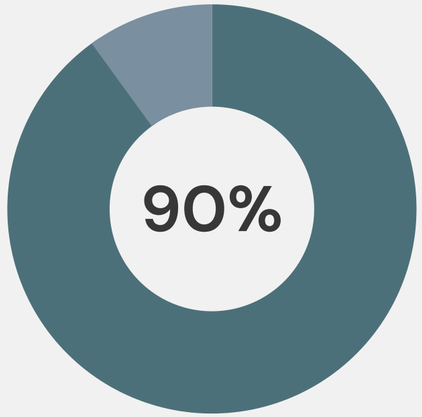
Model overall performance score is 0.99.



Precision

99% of predicted removed records are correctly predicted.

A high precision scoring that correctly classifies the correct class.



Recall

91% of actual removed records are correctly predicted.

Summary

A quick recap of our project

PPP Removed Application

Q&A

Appendix

Correlation

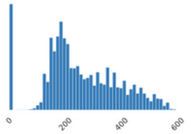
status_gen_cycle

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	566
Distinct (%)	0.2%
Missing	275744
Missing (%)	47.6%
Infinite	0
Infinite (%)	0.0%
Mean	255.2427514

Minimum	0
Maximum	593
Zeros	23666
Zeros (%)	4.1%
Negative	0
Negative (%)	0.0%
Memory size	4.4 MiB

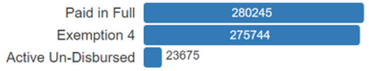


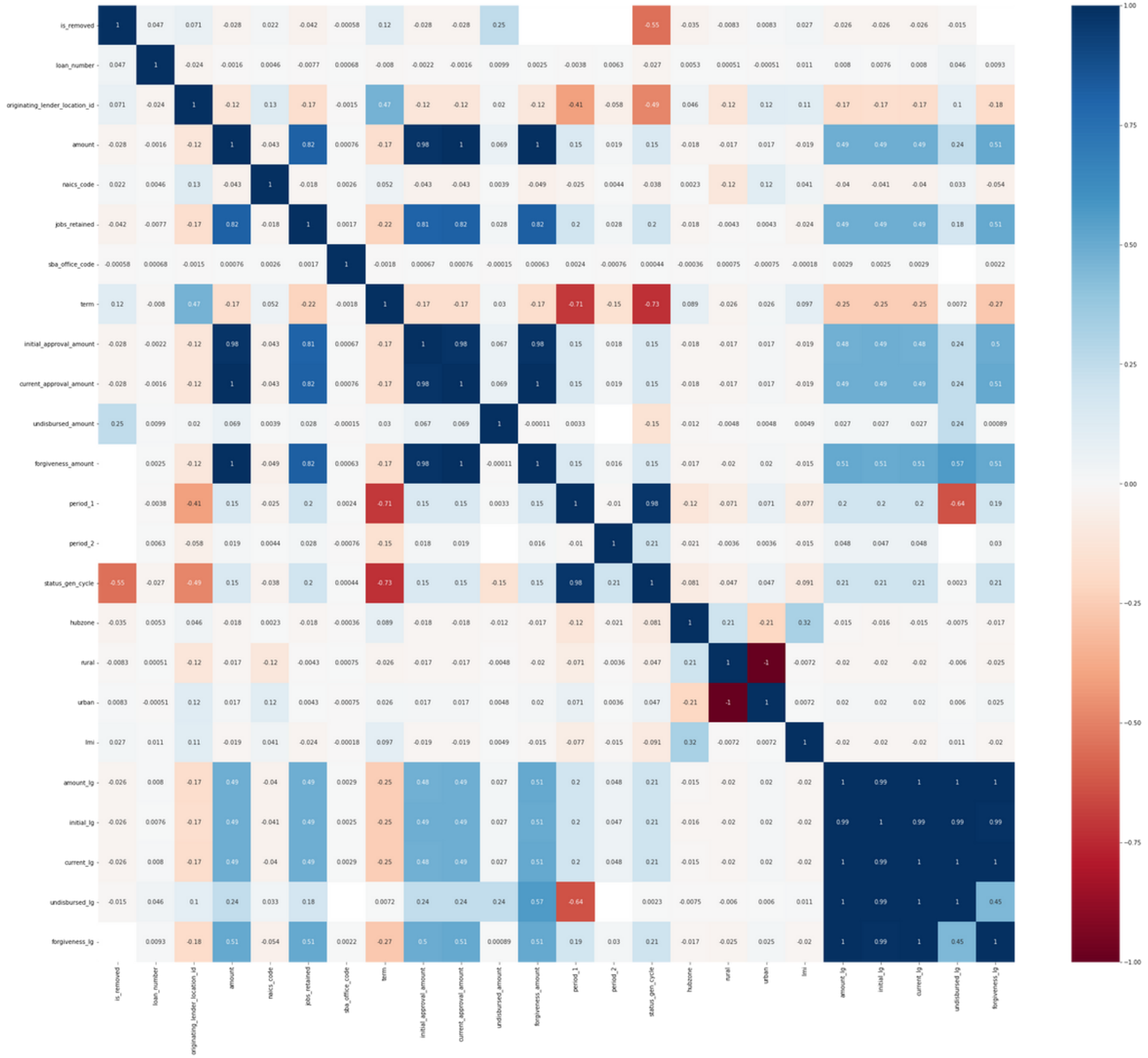
loan_status

Categorical

HIGH CORRELATION
HIGH CORRELATION

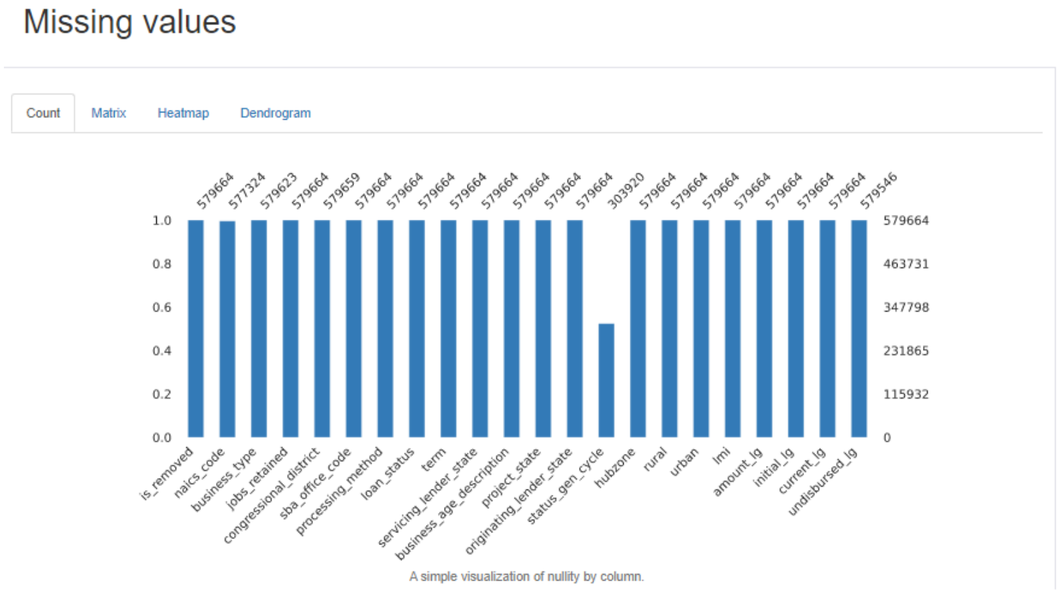
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.4 MiB





Appendix

Missing Variables



Appendix

Statistical Modeling

```
Optimization terminated successfully.
         Current function value: 0.005889
         Iterations 18
```

Logit Regression Results

Dep. Variable:	is_removed	**No. Observations:**	303920
Model:	Logit	**Df Residuals:**	303915
Method:	MLE	**Df Model:**	4
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.9783
Time:	22:26:51	**Log-Likelihood:**	-1789.8
converged:	True	**LL-Null:**	-82302.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.7165	0.149	31.646	0.000	4.424	5.009
status_gen_cycle	-0.1872	0.014	-13.357	0.000	-0.215	-0.160
jobs_retained	0.0255	0.010	2.428	0.015	0.005	0.046
hubzone	-0.5786	0.115	-5.025	0.000	-0.804	-0.353
urban	-0.4353	0.151	-2.880	0.004	-0.731	-0.139

Possibly complete quasi-separation: A fraction 0.92 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

```
Optimization terminated successfully.
         Current function value: 0.155117
         Iterations 11
```

Logit Regression Results

Dep. Variable:	is_removed	**No. Observations:**	579664
Model:	Logit	**Df Residuals:**	579659
Method:	MLE	**Df Model:**	4
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.1487
Time:	22:44:37	**Log-Likelihood:**	-89916.
converged:	True	**LL-Null:**	-1.0562e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.6749	0.018	-91.360	0.000	-1.711	-1.639
exemption4	-2.4646	0.021	-115.616	0.000	-2.506	-2.423
jobs_retained	-0.4075	0.007	-60.301	0.000	-0.421	-0.394
hubzone	-0.3443	0.015	-22.283	0.000	-0.375	-0.314
urban	0.1153	0.016	7.009	0.000	0.083	0.147

	2.5%	97.5%	OR
Intercept	83.463612	149.701131	111.779234
status_gen_cycle	0.806796	0.852362	0.829266
jobs_retained	1.004927	1.047110	1.025802
hubzone	0.447429	0.702652	0.560702
urban	0.481202	0.870185	0.647097

Appendix

Machine Learning

```
1 # Looking at the distribution of target variable
2
3 lgbm_df['is_removed'].value_counts()
```

```
0    553828
1     25836
Name: is_removed, dtype: int64
```

```
LightGBM Model training-set accuracy score: 0.9957
LightGBM Model testing-set accuracy score: 0.9953
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	221420
1	0.99	0.91	0.95	10446
accuracy			1.00	231866
macro avg	0.99	0.95	0.97	231866
weighted avg	1.00	1.00	1.00	231866