Loan Approval

Authors: Jing Zhang (wz223), Qiwen Li (ql257), Yuqing Xu (yx486)

Table of Contents

1. Abstract

Credit scores help banks, loaners, and others justify or evaluate a person's ability to return their debt, enabling them to make the right decisions in loan approval and helping disadvantaged individuals/groups to get the loan approved. Loan evaluation would help in minimizing loan risk, maximizing banks' profits, and protecting margins. Moreover, machine learning models can provide a more efficient and standardized way of loan approvals with less human intervention. Our project aims to evaluate the fairness of the current loan system at Bank of America across the United States in 2018 with data provided by HMDA Data Browser. Another goal is to implement a new system that is fairer compared to the original system.

2. Introduction to Project Dataset

The dataset that we are planning to use is a 2018 HMDA Dataset across the United States, and the financial institution would be Bank of America, National Association - B4 TYDE B6G KMZ 031 MB27. It contains 99 features, including ethnicity, race, loan type, income, action taken, preapproval, debt to income ratio, and so on, which is sufficient for us to find the factors that contribute to the current system. It also contains 377468 samples. Thus, this data set will allow us to answer this question since it provides a large amount of data, which would be enough for us to divide them into a training set, a validation set, and a testing set that would help us develop our new scheme and test the accuracy. We found the dataset on the FFIEC website.

2.1 Avoid overfitting/underfitting

Overall, to avoid overfitting/underfitting, we split the dataset into the training and testing set. We will apply validation methods such as examining accuracy in training and testing data while choosing parameters of models based on the validation set. Considering that we have many features, we are less concerned with underfitting. By looking at loan applications from all over the country, we can obtain a large sample size, and a large number of features would increase the model's complexity, which prevents overfitting. Furthermore, we have selected certain models and methodologies in our Exploratory Analysis to prevent overfitting.

3. Data Cleaning and Transformation

When we first examined our dataset, among the 99 features, we decided to first manually remove 63 features that we considered as irrelevant or duplicate information based on the feature definition and possible values of each feature presented on the data field website.

Next, we convert the data type for categorical features into string type values to input them into different algorithms in the future. Then, we treat categorical features that contain only two possible values as Boolean values corresponding to the values of 0 and 1. For categorical features with multiple values, we applied one-hot encoding. Furthermore, we decided to drop some of the outliers that we have identified in our data exploratory phase and looked through the columns with missing values.

Columns with Missing Values

```
conforming_loan_limit            94
loan_term                         1
property_value                 2429
income                        11138
debt_to_income_ratio           4808
applicant_race-1                 27
co-applicant_age_above_62    210312
dtype: int64
```

Figure 1.1

We decided to drop some of the insignificant rows that are missing in value in various columns, like `conforming_loan_limit, loan_term, property_value, debt_to_income_ratio, and applicant_race-1`, since there are less than 2% of our data don't have these values. For missing values in the applicant_age_above_62 column, we considered them as "not applicable" because the other two values are 1(yes) and 2(no) so we replaced the missing ones with value 3, indicating "not applicable". Lastly, we realized that the income column is also partially absent. So we created another feature, `missing_income,` that states if the value in the income column is missing which we thought would be interesting to examine later on. In order to fix the problem of missing income, we looked at correlations between other variables and income and we found out that loan amount is the highest correlated variable. With the loan amount, we applied the linear regression model to predict the missing income values based on the loan amount.

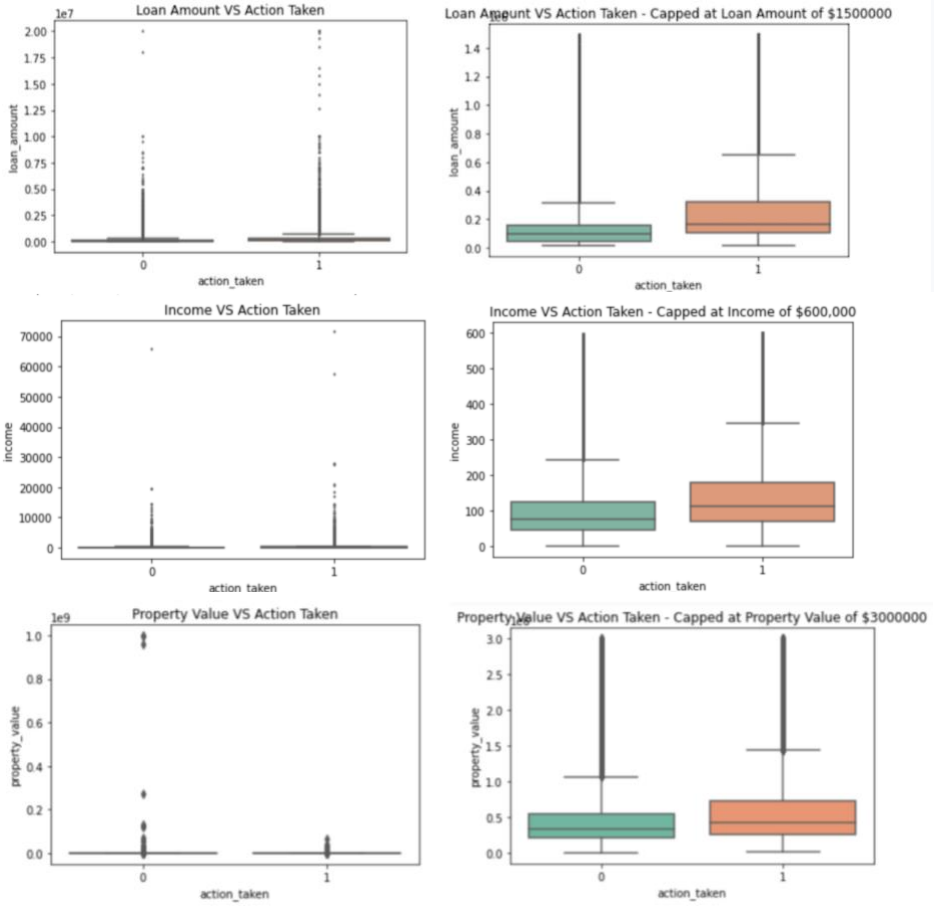
For our target feature, action taken, which tells whether the loan application is approved or not, we transform the feature value to 1 if it's approved, and 0 otherwise.

3.1 Number of features and Sample after Cleaning

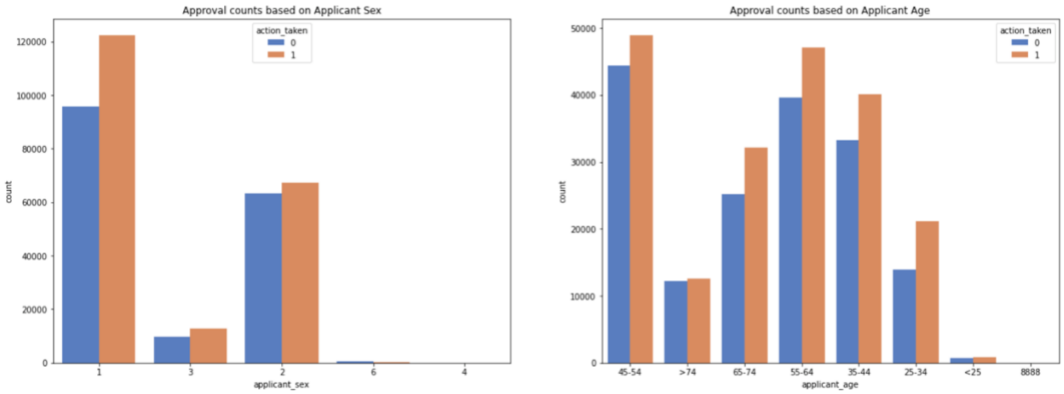
Now, the dataset contains documentation of 371,930 samples of loan applications with 31 features that we determined as valuable and insightful in determining our loan approval models. Some examples of the features are `derived_race, debt_to_income_ratio, loan_amount,` and `income`.

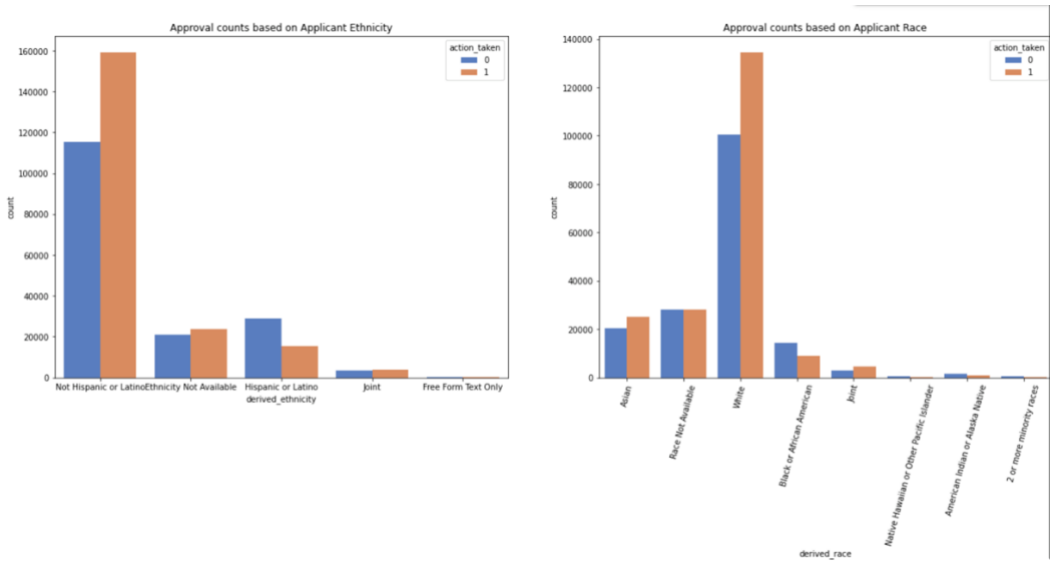
4. Data Visualization

To better understand our dataset, we created data visualizations regarding some of the relevant features that could be insightful. We constructed boxplots to see the spread of the data based on different features by loan approval status with the percentile distribution; it helps us identify outliers that were removed in later data-cleaning processes.



The above Box Plot confirms the presence of outliers/extreme values, which indicates the loan amount, income, and property_value disparities among all the loan applications. Furthermore, we would like to investigate if there are biases hidden in the loan approval algorithms by plotting histograms to see the variability of the numbers of loan applications that are approved or disapproved based on some features that we considered as biases which are age, race, sex, and ethnicity.





- The first histogram shows that the proportion of male applicants who got approved is higher than the proportion of female applicants who got approved, which indicates that the current system might have a gender bias.
- The second histogram conveys that the current loan approval system favors younger customers more than those older, especially those whose ages are greater than 74, which indicates that age discrimination might exist in the current system.
- The third histogram shows that the system favors Not Hispanic or Latino since the proportion of Not Hispanic or Latino applications who got approved is significantly higher than that of others, which tells us that ethnicity discrimination is not unlikely to happen in the current loan approval system.
- From the fourth histogram, we spot a large difference between numbers of loan approved and disapproved for white applicants, while small differences for other races. It shows that racism might be a factor in the loan approval system.

Overall, the number of applications approved is lower for underrepresented groups among each classification of sex, age, race, and ethnicity. According to the histograms, the loan approval system appears to favor male, younger customers, and people who are not Hispanic or Latino.

5. Exploratory Data Analysis (preliminary analyses)

5.1 Analysis Overview

We are exploring the most correlated features associated with the target feature and investigating whether those features are significant by performing regression (linear regression and logistic regression), and random forest. We split the features into grouped features with real-valued, Booleans, and categorical features. After data preprocessing, we divided the dataset into two parts, with 80% of the data as the training set and the rest 20% as the test set. We used linear regression for features with numerical values, and we used logistic regression on features with categorical values and Boolean values. We performed random forest to further investigate the top 15 important features of loan approval, white, and younger customer base.

5.2 Linear Regression

A model that we have implemented is Regression, we applied linear regression to real value variables and logistic regression to categorical and Boolean variables. Overall, Linear Regression and Logistic Regression works in a similar way. Linear regression uses a linear function to predict y that minimizes the sum of square residuals and returns numerical values associated with the target feature. By running a linear model on features with real-valued data, we can see that the `total_units` variable has the most significant impact on the loan approval process since it has the highest coefficient.

OLS Regression Results

Dep. Variable:	action_taken	**R-squared (uncentered):**	0.563
Model:	OLS	**Adj. R-squared (uncentered):**	0.563
Method:	Least Squares	**F-statistic:**	6.854e+04
Date:	Sat, 30 Oct 2021	**Prob (F-statistic):**	0.00
Time:	21:05:50	**Log-Likelihood:**	-2.6087e+05
No. Observations:	371930	**AIC:**	5.218e+05
Df Residuals:	371923	**BIC:**	5.218e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
loan_amount	3.09e-07	2.99e-09	103.411	0.000	3.03e-07	3.15e-07
loan_term	0.0007	1.18e-05	55.029	0.000	0.001	0.001
property_value	-2.972e-09	2.77e-10	-10.728	0.000	-3.51e-09	-2.43e-09
total_units	0.1404	0.003	41.210	0.000	0.134	0.147
income	-3.757e-05	3.67e-06	-10.224	0.000	-4.48e-05	-3.04e-05
tract_minority_population_percent	-0.0013	3.16e-05	-40.292	0.000	-0.001	-0.001
tract_to_msa_income_percentage	0.0012	1.73e-05	69.121	0.000	0.001	0.001

Omnibus:	3266480.971	**Durbin-Watson:**	1.422
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	29267.926
Skew:	-0.191	**Prob(JB):**	0.00
Kurtosis:	1.680	**Cond. No.**	1.27e+07

5.3 Logistic Regression

We also applied a logistics model on categorical and Boolean features. After obtaining the model summary from each feature, we decided to take features with coefficient < 0.01 out of consideration regarding what factors can potentially influence the loan approval status. By comparing the coefficients from the summaries of different features, we found that `missing_income, applicant_sex, applicant_age_above_62` are the important features among all the input features for the logistic model.

6. New Model Building

6.1 Logistic Regression Classifier (Attempt)

Logistic regression is used when the target variable is a categorical variable, and it made sense in our analysis since loan approval is also categorical. We built a logistic regression classifier model using the training set, and we found that the testing accuracy for such a model is quite low, so we found another way to improve on the logistics model, which is using the grid search.

```
Tranning Accuracy of Logistic Regression Classifier:  0.6490871938267954
Testing Accuracy of Logistic Regression Classifier:  0.627040034415078
```

6.2 Grid Search (Attempt)

We performed a grid search and found that the test accuracy improved by nearly .03, since this improvement is not significant, we decided to build another model: random forest.

```
Best Score: 0.6508583431464181
Best_Parameters (0.6508583431464181, {'max_iter': 100, 'tol': 5})
```

6.3 Random Forest (Final Model)

Another model that we have implemented is Random Forest. Random Forest performs a bootstrap aggregated ensemble model of trees containing random subsets of features and it can be applied to both regression and classification. The benefits of Random Forest include reduce overfitting, improving accuracy, efficiency in handling large datasets, and others. The most important attribute of Random Forest is its ability to rank feature importance which would provide massive insight toward our research question to determine the key predictive features that are used to determine loan approval in the current BOA loan approval process. With the current models that we have constructed so far, we are able to determine the features with stronger significance in predicting loan approval. We will then perform cross-validation to prevent overfitting and selection bias.

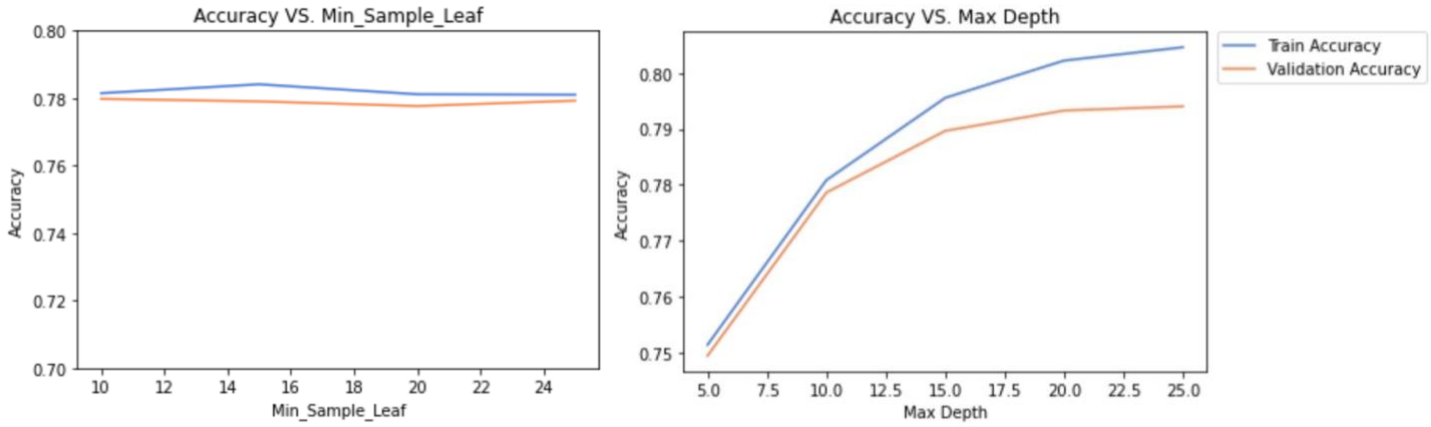
6.3.1 Hyperparameter tuning

```
min_samples_leaf=list(np.linspace(15, 40, num=5).astype(int))
w=cross_validation(train_x_clean,train_y_clean,val_x,val_y,min_samples_leaf)
```

```
w
```

```
(15, array([0.27039845, 0.27116024, 0.27234774, 0.27290041, 0.27530528]))
```

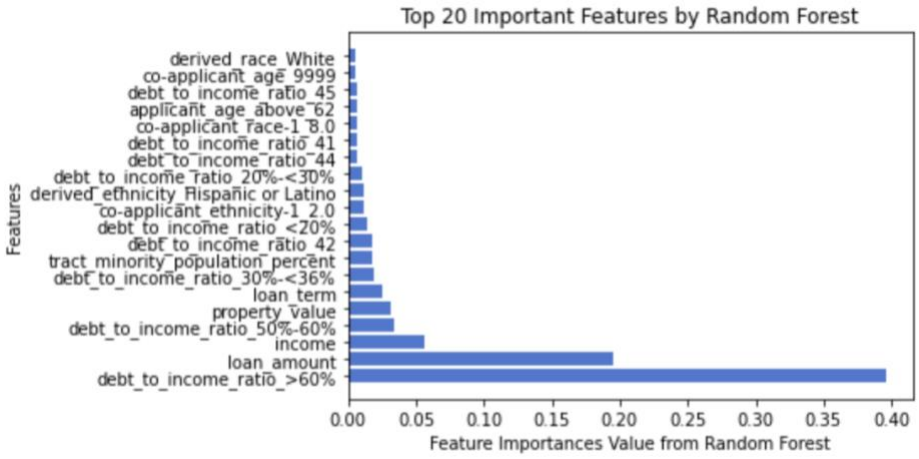
A strategy that we implemented to prevent overfitting and to optimize our model is cross-validation. Cross-validation allows us to find the best parameters that optimize the accuracy of the model by performing a resampling procedure repeatedly. We utilized 10-fold cross-validation to execute hyperparameter tuning for our Random Forest model with a validation set that is 20 percent of the training set. We held the parameter, n_estimators, to be a constant of 100 due to the runtime, then looping through the possible values of min_samples_leaf between 15 to 40 and max_dept between 5 to 25. We made a deliberate decision in setting the lower bound of our min_samples_leaf parameter to be 15 to prevent overfitting/underfitting and we chose the upper bound of 40 in order to still examine a wide range of options. Same decisions are made for the lower and upper bound of parameter max_dept to avoid overfitting/underfitting. We trained the model with the hyperparameters and evaluated it based on how well it scored on the validation data. By examining the first line graph below, "Accuracy VS. Min_Sample_Leaf", we see a peak at min_sample_leaf equals 15 which corresponds to the highest train accuracy and validation accuracy. The line graph "Accuracy VS. Max. Depth" suggests that the optimal max_depth is around 15 since the margin between train accuracy and validation accuracy is within 0.01. In the end, we concluded that the minimum error amount occurred at the optimal parameters of min_samples_leaf equal 15 and max_depth equal 15.



Moreover, we have attempted to perform boosting to increase model complexity and accuracy, but we discovered that it overfits the training model resulting in lower test accuracy. We decided to then discard that implementation, the attempt is still available in our Python Notebook.

6.3.2 Result

The result of the random forest model is described in the bar graph below with the printout of the Top 20 Important Features by Random Forest. It illustrates that `debt_to_income_ratio`, `loan_amount`, and income are the most important factors toward loan approval. However, `applicant_age_above_62, co-appliacnt_ethinicity, co-applicant_race, derived_race, tract_minority_population_percent,` and `derived_ethnicity` also play a role in determining whether a loan should be approved, indicating that the bank does take unfair elements into consideration and their loan approval system does appear as biased. We considered these features as unfair because they are the factors that people cannot physically change, they are identities that people are born with or have no control over.



Looking at our accuracy values for training and testing sets, we obtained relatively high accuracy for both subsets of data. This means we are able to avoid overfitting and the model does not underfit because the model is complex, and we used feature selection to select highly correlated features then implemented cross-validation to improve our accuracy. Our training accuracy is 0.804, validation accuracy is 0.794, and the testing accuracy is 0.801.

```
Train F1 Score  0.8264195193961796
Train Accuracy  0.8039886537789369
Validation Mean F1 Score:  0.7903971487513262
Validation Mean Accuracy:  0.7941279257187839

Accuracy of RandomForestClassifier:  0.8011050466485629
```

To create a better model, hence a model that is fairer, we decided to remove all the unfair features that was listed above (`applicant_age_above_62, co-appliacnt_ethinicity, co-applicant_race, derived_race, tract_minority_population_percent, and derived_ethnicity`). By doing so and training it again on the Random Forest Classifier, we achieve a training accuracy of 0.7966 and a testing accuracy of 0.803716 To conclude, although our training accuracy did decrease after removing the unfair features and our testing accuracy is around the same, we do see a slight improvement.

```
Training Accuracy of RandomForestClassifier:  0.7966163215099514
Testing Accuracy of RandomForestClassifier:  0.8037157030126232
```

7. Fairness

Fairness is the most important criterion to consider when choosing a model since that is the goal of our project. Examining the top 20 important features of the current model we see some variables violating the fairness of Equality of Opportunity where features relating to race, color, national origin, age, or other are considered in the loan application process. We also realized problems in Unawareness where there exist covariates that are correlated to the protected attribute, for example, the `tract_minority_population_percent` feature is indirectly related to racial groups. On the other hand, we understand that Demographic Parity would make sense in terms of features such as age. People within the age group of 25 to 40 are more likely to take out loans and at the same time they are more likely to be approved for loans since they are more likely to obtain a stable job with a stable income while compared to an applicant who is 15 years old and still in school. Predictive Rate Parity also makes sense since the bank would want to know what fraction of people that they approve for loan actually pay back the loan, it encourages equal error reduction in all groups. Even though, it's a conditional fairness because you would never be able to measure the pay back rate for people that are not approved. But at the end, the bank is aiming to minimize the risk.

8. Conclusion

By using techniques that we learned in class, such as feature selection, linear and logistic regression, cross-validation, logistic regression classifier, grid search, and random forest, we investigated the fairness of the current loan approval system and found out some features that are biased. We used validation sets, k-fold-validation specifically to test effectiveness. We also plan to compare MSE, p-values, and R-squared values for different models to choose the optimal model. Additionally, we built a new model with all the biased features removed and found out that training accuracy and testing accuracy are both approximately 0.80, which indicates that it's a less biased model.

Furthermore, we constructed the Confusion Matrix of the testing dataset, to examine our prediction errors and Equalized Odds. In which, we do see that the Random Forest model is predicting much more false negative than false positive, which means a lot more applications are being approved although they should not be. This issue should be addressed in further implementation, where we will rather have more false positives than false negatives (i.e., rather predicting wrong for those who should be approved than predicting wrong for those who should not be approved.)

Confusion Matrix (Test Data)

Predicted_Y_test / Y_test_true	0	1	All
0	19660	9418	29078
1	5377	39931	45308
All	25037	49349	74386

9. Appendix

- Data cleaning process: https://colab.research.google.com/drive/1OyOJC7SnGlmoUiwuDzbzjDWORMNzCqSP?usp=sharing
- Cleaned dataset: https://drive.google.com/file/d/1-5cKDMIyaZYH2Z3h4-9qzNv5qimxIq4b/view?usp=sharing

10. References

- Original Dataset: https://ffiec.cfpb.gov/data-browser/data/2020?category=nationwide&leis=B4TYDEB6GKMZO031MB27
- Data field website: https://ffiec.cfpb.gov/documentation/2018/lar-data-fields/
- *Assessing and Mitigating Unfairness in ... - Microsoft.com*. https://www.microsoft.com/en-us/research/uploads/prod/2020/09/Fairlearn-EY_WhitePaper-2020-09-22.pdf.