Fairness in Lending
Mike Labadie

Background

Property ownership has been a source of racial inequity in this country since its founding. With progress in civil rights, legislation has been enacted to prohibit and monitor discriminatory practices in the real estate lending market.

The Importance of Fair Lending

Lending helps provide access to housing with close proximity to good jobs, vital amenities like public transit, and good schools for development of younger generations. Home ownership is also a method for wealth accumulation and the passing wealth on to future generations.[6]

The United States has long had a wide wealth gap between races despite improvements in civil rights. This problem has been long perpetuated through discriminatory practices in home ownership. One such practice, known as redlining, saw home mortgage lenders exclude neighborhoods comprised predominately of minorities.

Legislation

The Fair Housing Act, which was enacted in 1968 as part of the Civil Rights Act, prohibited discrimination in the sale, rental, and financing of residential dwellings. In 1975, Congress enacted the Home Mortgage Disclosure Act[2] (HMDA) to increase transparency into lending practices. The HMDA "requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages."[1] The disclosure of lending activity is currently required of institutions that originated 25 mortgages or more in a given year. Under the new "Economic Growth, Regulatory Relief and Consumer Protection Act," signed by President Donald Trump on May 24, smaller banks are no longer are required to report detailed information about mortgage loan applicants.[4] This new legislation raises the cut off to institutions that originated 500 or fewer loans.

Home Mortgage Disclosure Data for 2017

HMDA data is made freely available online at the Consumer Financial Protection Bureau's (CFPB) website[8]. The 2017 file captured over 14 million loan applications from across the United States. These loan applications were made with ~5,700 financial institutions. The file captured a variety of fields related to the specific borrower, as well as characteristics of the census tract of the home location. This change could dramatically change the nature of the HMDA dataset. Based on 2017 data, this would exempt almost 4,300 lenders from reporting (nearly 75%). However, the exemption of these smaller lenders would reduce the loan application count by just 7%.

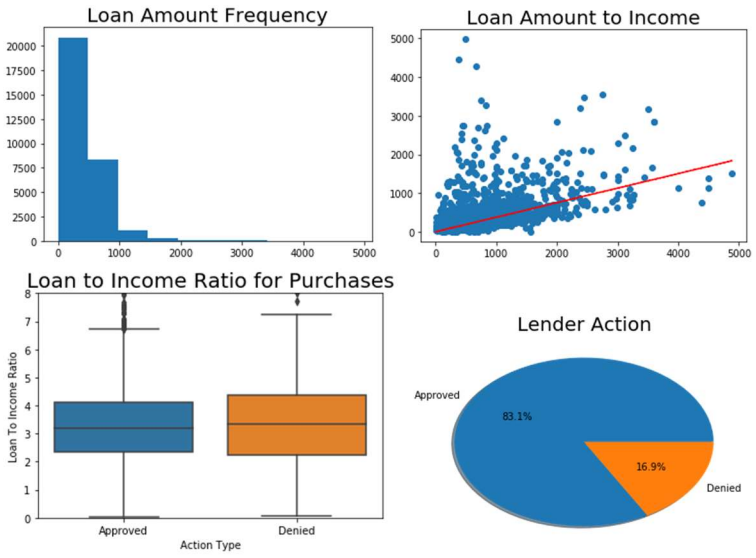
Exploratory Data Analysis

Findings

Before beginning the modeling exercises, I undertook an iterative exploratory data analysis process. The goal of this process was to attain a better understanding of the characteristics of the data. This understanding boosted the quality of the modeling process. I first created a comma-delimited file that contained only loan applications for the District of Columbia. While being directly relevant to me, I concluded that D.C. would provide a good base to study given that it is a large and diverse metropolitan area. The persistence of this comma-delimited file aided development as it didn't require each exercise to filter the large file; while also centralizing any field mapping or calculated field creation. The code used to create the D.C. file can be found in the "1. Build DC File" notebook. The D.C. file contained nearly 31k loan applications; with nearly 50 characteristics for each application. This exploration also highlighted the varying levels to which each lender populated values for each characteristic; ranging from complete coverage to extremely sparse coverage. This analysis can be seen in the "2. Descriptive Basics" notebook.

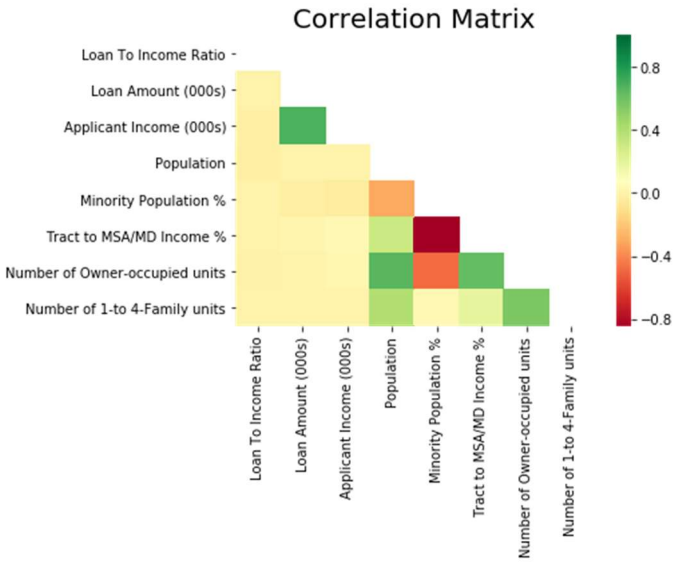
Robust analysis of the data is found in the "3. EDA" notebook. As seen below, most loan applications fell below $1 million. Loan amounts showed a positive relationship; though with a large amount of variability. Loan

Amount to Income ratio was slightly higher for applications that were denied. This ratio showed a wider range of variability than approvals. Finally, as the pie chart shows, over 80% of loan applications were approved.



Correlation

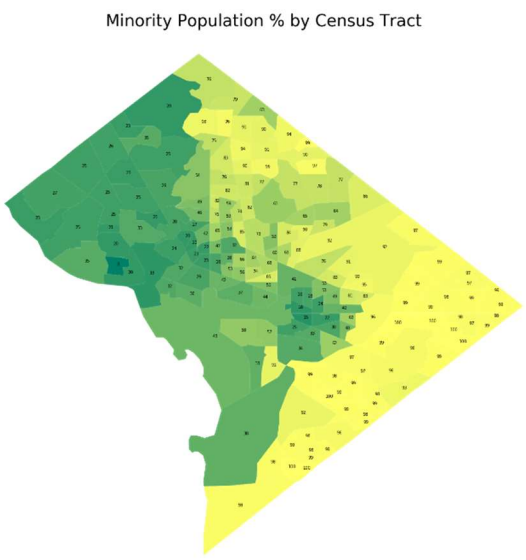
From the correlation matrix below, you can see the strong positive relationship with loan amount and income. It is intuitive that higher income earners can pay back larger loans. It is also intuitive to see the positive relationship between population and the number of housing units. The larger population requires a higher number of living units. The correlation matrix also illustrates two unfortunate relationships. There is a strong inverse relationship between the minority population percentage and the income level of a given census tract. Also, there is a moderately negative relationship with the number of owner-occupied units and minority population percentage of a given census tract. These discoveries confirm the social inequities still present in income and home ownership.

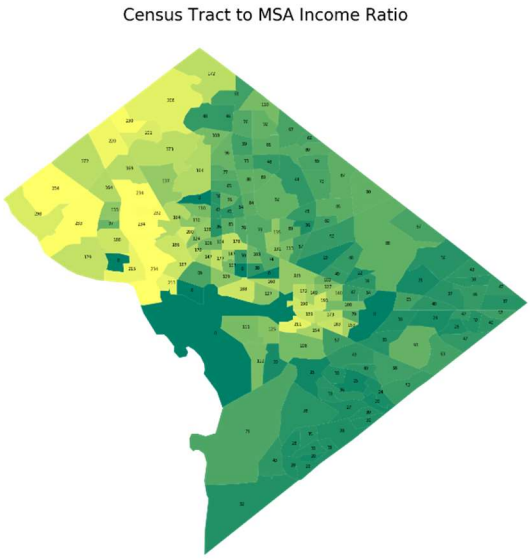


Geographic Visualization of the Inverse Relationship of Minority Population and Income

The images below show relationship between minority population percentage and the income level at the census tract level for the District of Columbia. The lighter shaded region on the first image show the higher

percentage of minority population of the census tracts in the southeast. The darker region on the second image show the lower income levels of those same census tracts.





Modeling

The goal of the subsequent modeling exercise is to classify the loan application records on the action taken by the lender; meaning whether the loan was approved or denied by the lender. While being an informative study of various modeling techniques, I'm also seeking to highlight where each technique can help identify loan application characteristics that have the most importance in determining a lender's action.

Modeling: Random Forest

Initial Model

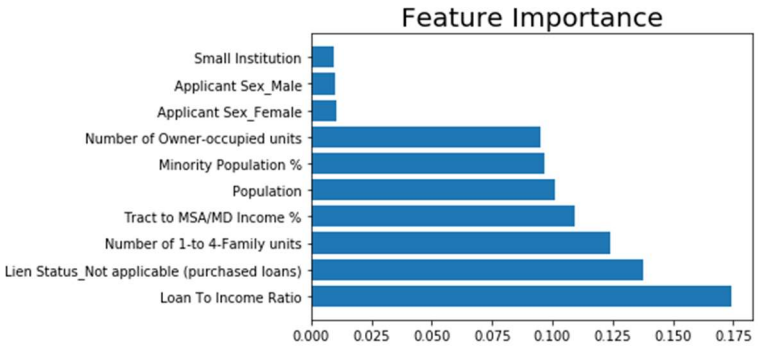
The initial, untuned random forest model exhibited an overall accuracy of 81%; roughly the percentage of the approved applications observed in the dataset. However, the "denials" accuracy was just 35%. In this case, accuracy is largely based on predicting a high number of approvals. Unfortunately, that doesn't present a profoundly insightful, nor powerfully predictive, model.

Tuning

I first adjusted the number of trees generated through the random forest process (i.e. n_estimators). Accuracy changes very little with n_estimators set to 10,000 (vs the default of 10). Interestingly, n_estimators set to 3 consistently achieved higher "denial" accuracy (+2-6%). Conversely, n_estimators set to 2 saw significantly lower "denial" accuracy (as much as -13%). Next, I needed to properly handle the unbalanced nature of the classification classes. Adjusting the class_weight parameter resulted in higher accuracy of predicted denials.

Feature Importance

When determining whether to extend a loan, lenders will ask two very basic questions: Can the borrower pay me back? And if they do not pay me back, how can I recover what I paid for? Hence, one would expect income ratio and lien status to be two important features in predicting lender actions. It is comforting to see the model found Loan to Income as the most important feature, and for the importance to be so much higher than the next feature. However, it is concerning that Minority Population percentage for the census tract is one of the more prominent features. Finally, applications to lenders exempted from reporting by the newly enacted legislation was the 10th most important feature.



Modeling: Logistic Regression

Initial Model

My initial logistic regression model showed an 80% accuracy classifying whether a loan would be approved or denied. However, this was a worthless model. This model classified every application as "Approved". As with the random forest model, this was the result of the bias caused by having an imbalance in the distribution of approvals and denials.

Tuning

The solution is to introduce a mechanism that encourages the model to give a higher predicted probability to the rare class. Adjusting class weights, penalizes the model for misclassifying the minority class.[7] The class_weight parameter for the final model was set to "balanced".

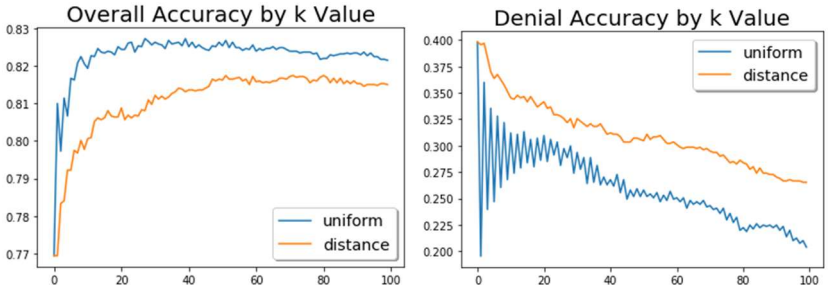
Modeling: K-Nearest Neighbors

Brief Overview

K-Nearest Neighbors (kNN) is a simple classification algorithm that matches an unseen sample with the k most similar previously seen samples. A function based on all input features determines the similarities ("distance") between samples. The majority classification found in the k nearest samples is the predicted classification. My KNN model can be found in the "6. k-Nearest Neighbors" notebook.

Tuning

Even values for k are unreliable when you have a binary classifier because the model randomly breaks ties. See the "Denial Accuracy by k Value" chart below. Overall accuracy rises moderately. kNN is prone to same bias errors as other models resulting from imbalance. kNN is more prone to sample size imbalance as k rises. Attempt to correct this with distance weighting function, where further neighbors are given less weight in each sample's prediction.



Summary

Results for Each Technique

k-Nearest Neighbors

Actual \ Predicted	Approved	Denied
Approved	3005	354
Denied	513	301

Overall Accuracy: 79%
Denials Accuracy: 37%

Random Forest

Actual \ Predicted	Approved	Denied
Approved	2615	744
Denied	277	537

Overall Accuracy: 76%
Denials Accuracy: 66%

Logistic Regression

Actual \ Predicted	Approved	Denied
Approved	2435	924
Denied	187	627

Overall Accuracy: 73%
Denials Accuracy: 77%

Conclusion

There are a number of methods for evaluating the quality of the models: overall accuracy, rare classification accuracy, interpretability. While achieving higher overall accuracy may be appealing, there could be instances where understanding and predicting a certain class may be a higher priority; as we have with this exercise of understanding potentially unfair lender practices.

All models initially struggled to handle the imbalance of the lenders' actions, which exhibited an ~80/20 split. This warranted the introduction of mechanisms that would encourage the model to predict more of the rarer ("Denied") classification, or rather, choose less of the more common class. Performance of the Random Forest and Logistic Regression models improved dramatically when class weighting adjustments were introduced.

In their final forms, the kNN model had the highest overall accuracy, but poor performance in identifying denials. Both the Random Forest and Logistic Regression models achieved a much higher accuracy predicting denials, while suffering only a slight degradation in overall accuracy.

References:

- 1: https://www.consumerfinance.gov/data-research/hmda/learn-more
- 2: https://s3.amazonaws.com/files.consumerfinance.gov/f/documents/bcfp_hmda_2017-mortgage-market-activity-trends_report.pdf
- 3: https://www.ffiec.gov/ratespread/newcalc.aspx
- 4: https://pilotonline.com/business/banking/article_68862681-6352-5e4f-a577-02b3d8781e18.html
- 5: https://www.charlestonchronicle.net/2018/06/05/redlining-settlement-fails-to-provide-strong-penalties-other-actions-signal-more-backward-turns-on-fair-housing/
- 6: https://www.forbes.com/sites/lawrenceyun/2016/08/12/why-homeownership-matters/#4445bf7a480f
- 7: https://www.linkedin.com/pulse/welcome-real-world-data-imbalance-ian-clark
- 8: https://s3.amazonaws.com/cfpb-hmda-public/prod/snapshot-data/2017_public_lar_csv.zip