Racial Bias in Conventional Mortgage Pricing and its Relation to Solar Installation

As the Coronavirus ravages our nation, recent events have shed light on the more virulent strain of racial inequity that has plagued modern society. To help combat this disease, I decided to use my data science skillset to help contribute to the discussions and conduct a research project on identifying racial bias in conventional mortgages and its relation to Solar PV installations as part of the Data Good student club. Specifically, the theory I wanted to test was the idea that in some Metro Areas, Racial discrimination in Conventional Mortgages has resulted in minority groups obtaining an unfairly high monthly payment for their loans compared to whites. Metro Areas with this racial bias will have a wider solar installation gap between minorities and whites than MSAs (Metropolitan Statistical Area) without bias even when accounting for differences in homeownership to renter ratios. To test this hypothesis, I used the 2018 Home Mortgage Disclosure Act (HMDA) data, the 2018 Q1 Fannie Mae Single Family Performance Data, ACS 2018 5-year data, and Stanford's Deep Solar data which is updated as of 2018.

Part 1: Identifying Racial Bias in Conventional Mortgages

Background

(Source: The legal framework and background for this part are derived from a 2019 Berkeley Haas paper: https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf)

To identify whether minorities in some metro areas have received unfairly higher monthly loan payments due to racial bias, I will be trying to identify statistically significant interest rate gaps. Even if these gaps are small, these tiny differences add up over time due to monthly compounding. Yet, it is important to realize that not all interest rate gaps can be attributed to racial discrimination as any lender with a gap can justify the difference on the principle of a legitimate business necessity, which is defined as a lender's use of variables and practices to ascertain creditworthiness. To understand the implications of this, consider an example where the lender uses education level as a predictor of credit worthiness. Since there are racial differences in education, the use of this variable may create a racial interest rate gap, but legally, it would not be considered potential racial discrimination due to the legitimate business necessity defense. Thus, to identify racial bias, we need to capture the full scope of credit risk.

This task can be accomplished by only considering loans from the Government Sponsored Entitles (GSE) Fannie Mae and Freddie Mac. These institutions are part of the secondary mortgage market. They buy mortgages from lenders and package them into an asset called a Mortgage-Backed Security (MBS), which are then sold to investors.

When you apply for a GSE backed mortgage from your lender, what happens is that it first goes through the Fannie Mae/Freddie Mac underwriting system. Based on your loan to value ratio, and your credit score, the GSE assigns your loan a g-fee according to the Loan level Price Adjusting (LLPA) matrix. This is the extra amount the government adds on to your interest rate to cover the cost of your credit risk. From there, the lending institution will add on the risk-free rate for a loan and may add more to your interest rate to cover origination charges or for predatory reasons. The GSE will then buy the loan from the lender.

Therefore, limiting the scope to GSE loans removed the legitimate business necessity defense by providing a full picture of credit risk, which allows us to construct a regression model to identify racial interest rate gaps.

The Regression Model

For each MSA, I ran the following regression model for all the loans in that region:

$$interest\ rate = a * race + u_{sex} + u_{age} + u_{gfee} + u_{month} + u_{lender} + b1 * discount\ points + b2 * discount\ points * lender + b3 * origination\ charges + b4 * origination\ charges * lender + e$$

Here is what each variable in the model means:

Dependent Variable:

The dependent variable is interest rate in basis points. A basis point is 100* rate. All our rates are in basis points for this study.

Independent Variables:

$a * race$: Race is the race for the applicant- Black, Hispanic, Asian, White, or N/A. For analysis though we are interested in the coefficient(a) for the Black and Hispanic indicators. This value is taken relative to whites (reference variable) so it represents the interest rate gap we want to measure.

The rest of the variables in the equation is for factors we want to control.

$u_{sex} + u_{age}$: Represents demographic fixed effects and is the age and sex of the loan applicant. Age is not the exact age but instead the HMDA data provides age groups.

$u_{gfee} + u_{month} + u_{lender}$: G-fee is the value ascribed to the loan from the LLPA matrix. Month represents the estimated month of the loan lender is the lender fixed effect.

$b1 * discount\ points + b2 * discount\ points * lender$: This is how we account for discount points. Discount points allow you to lower your interest rate for an upfront fee. Lender credits allow you to raise your interest rate by accepting a lower upfront payment. I decided to combine these two fields in the HMDA by performing discount points-lender credits for each loan. The value discount points in the regression is the combined value. Furthermore, we add the interaction term because the impact of discount points on the interest rate depends on the lending institutions policies.

$b3 * origination\ charges + b4 * origination\ charges * lender$: Origination Charges are costs the lender endures to service your loan. This impacts interest rate, and like discount points, the impact will depend on the lending institution.

A positive and significant value of "a" shows racial bias since we control for applicant credit risk(g-fee), macro-economic environment(month), consumer choice (discount points), and business expenses (origination charge). Thus, any significant gap remaining must be evidence of potential bias.

Linear Regression Assumptions Checks:

In order to properly interpret the coefficients in the regressions, we need to ensure that all our regressions follow the linear regression assumptions. Here are the tests used to ensure this:

Residuals are Normally Distributed: Assert that the skew of the residuals is between -1.2 and 1.2. Used. skew () function

No Multicollinearity: Conducting the Variation Inflation Factor (VIF) of all variables. Since we are only interested in race, we ensure that the race indicators have a VIF of less than 10.

No Autocorrelation: Pass the Durban Watson Test (Ensure result from this test is between 1.5 and 2.5.)

Homoscedasticity: Asserted that the correlation coefficient between residuals and predicted values are zero.

Furthermore, I also made sure to only report a coefficient if the corresponding race had more than 50 loans.

Important Filtering: Only originated ,first -lien loans, Debt to income <60%, only Fannie Mae and Freddie Mac loans, only 30 year fixed rate, fully amortizing, non-commercial loans, loan amount>30000, loan to value ratio between 30 and 130. Also, I removed the loans from the top 15 lenders for each MSA. This is due to large lenders holding Mortgage Back Securities which could impact risk scoring. Removing the top 15 mitigates this and increases the robustness of the results.

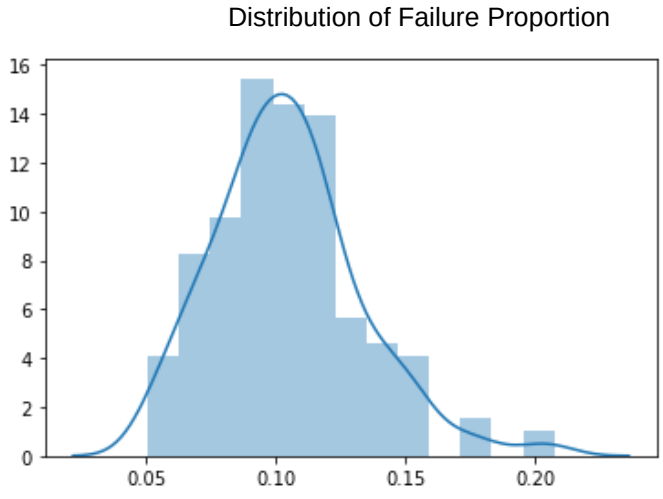
Estimating Month and Credit Score

Month and Credit Score are not part of the HMDA dataset. Thus, they need to be estimated.

To estimate month, I decided to calculate the approximate APOR for each loan. APOR is the prime-offer APR for a mortgage loan. The FFIEC provides a table with historic APOR values the date it took place.

Here are the steps taken to calculate the month for each loan:

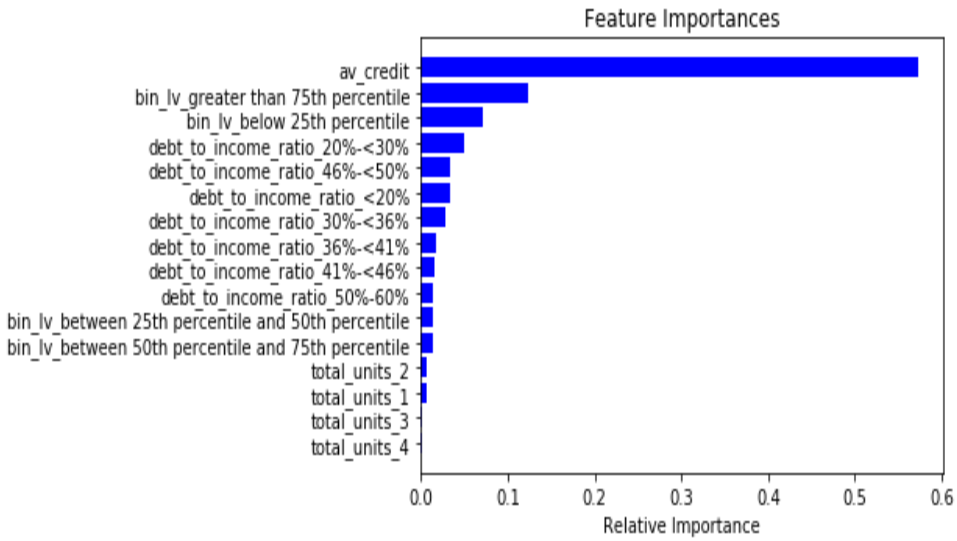
1. Calculate Cost of each loan using formula provided by the NCRC.
2. Find the Monthly Payment by using Mortgage library written by Austin McConnell
3. Create a list with the loan payment schedule. The first value is loan amount – cost, and the next 360 values are -monthly payment. Then use the internal rate of return (irr) solver (newton) written by erikbern. This is the APR of the loan.
4. APR-Rate Spread=APOR
5. Round to 2 decimal places and find Month with the nearest APOR in the FFIEC table



However, because the HMDA does not contain all the costs like Mortgage Insurance, this method did not work all the time. The failure proportion is shown in the histogram.

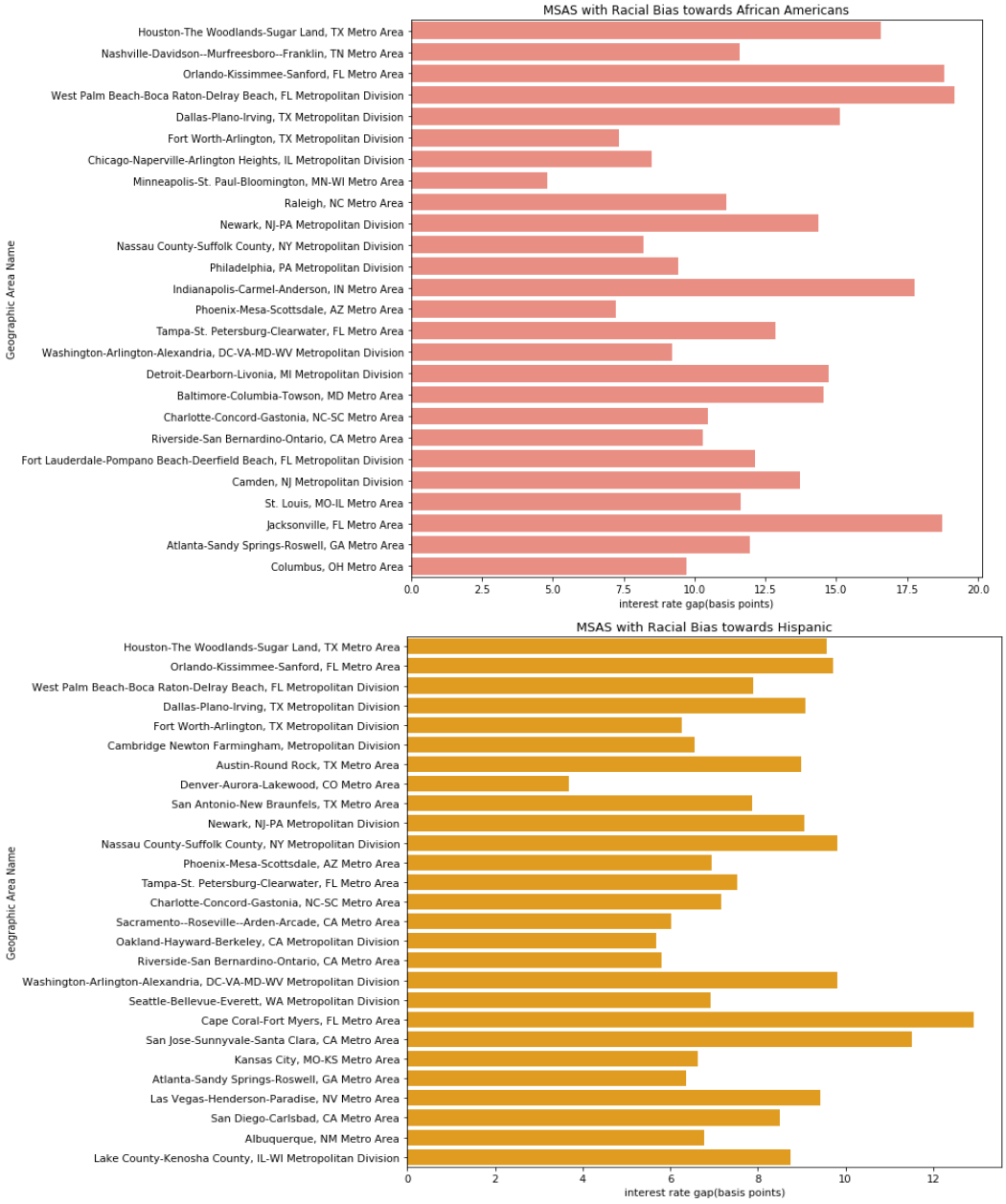
To estimate credit score, I decided to create a Machine Learning model using the Fannie Mae Q1 2018 Dataset. I took the features that were in both the HMDA and Fannie Mae data: Loan to Value ratio, Debt to Income Ratio, and the total units of the house. Using the Fannie Mae data, I also added the average credit score for the MSA in the model. Furthermore, I decided to bin the debt-to-income variable to match the HMDA classification, and binned the loan to value ratio by quartile.

After preprocessing, I tried two models to predict credit score- linear regression and Random Forrest. The Linear Model achieved a 0.11 r^2 and a RMSE of 40 on the test set. The Random Forrest did much better and got a 0.25 r^2 and a RMSE of 38 on the test set. Therefore, I decided to use the Random Forrest model to predict credit score. The bar chart shows the relative feature importance for each variable in our chosen model. As you can see, the Average Credit Score carries the most predictive power.



Results

I then ran the regression model for each Metro Area in the dataset. For African Americans, I was able to find 26 MSAs with significant interest rate gap and thus racial bias. For Hispanics, I was able to find 27 MSAs with significant interest rate gaps. Here are the Metro Areas:



Part 2 Association with Solar Installations

To understand whether this racial pricing bias had an impact on solar installations, I asked my partner at Data Good (Kelly Trinh) to perform two A/B tests- one for African Americans and one for Hispanics. The hypothesis we were testing was that MSAs with racial disparities against Hispanics/Blacks would have, on average, a higher solar installation gap between the minority group and whites than compared to MSAs without racial disparities. The null hypothesis is that any difference is due to random chance.

To calculate the Solar Gap for each MSA, I merged the data with the ACS 2018 census tract racial profile data, number of household's data, and homeowner to renter data. I took the census tracts that were

Majority white, black, and Hispanic, and were above the 50th percentile in homeowner to renter ratio. This is to minimize the renter effect as minority communities are shown to rent more and thus install fewer solar panels. I then calculated the portion of houses with solar for majority white, black and Hispanic census tracts. To find the gap, I simply subtracted the majority Black/Hispanic proportion from the majority white proportion.

We then ran the A/B tests to determine whether there was a significant difference between MSAs with racial discrimination and without. However, our p values were not lower than 0.05. Thus, we fail to reject the null hypothesis.

Conclusion

Even though we were unable to find a link with solar installation, this research was important as I was able to find metro areas with racial pricing bias across the United States in 2018. This shows that racial inequality is still alive in the Mortgage industry. For future research, it would be good to access data for more years and get exact values for the date and credit score.